INVESTIGATING THE SPACE-TIME VARIATION IN FINE PARTICULATE MATTER POLLUTION IN THE NORTHEASTERN UNITED STATES, 2000 – 2014

by
Stacy Elizabeth Woods

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
May, 2016

ABSTRACT

The health effects of fine particulate matter ($PM_{2.5}$) pollution exposure include increased risk of cardiovascular and respiratory illness and premature mortality. Since 2000, national levels of ambient $PM_{2.5}$ concentrations have decreased. The research presented in this dissertation document builds upon previous work to investigate $PM_{2.5}$ pollution trends in the northeastern United States (NE US) and identifies areas where people may be at greater risk of adverse health effects from $PM_{2.5}$ pollution exposure.

The objective of this dissertation is to characterize the variation in $PM_{2.5}$ over space and time in the NE US from 2000 to 2014. To accomplish this objective, we perform spatial analyses driven by three specific research aims. Aim 1 examines the dynamic relationship between environmental determinants and $PM_{2.5}$ pollution. Aim 2 assesses federal regulations designed to decrease $PM_{2.5}$ pollution and applies an innovative approach to evaluate the small scale variability of $PM_{2.5}$ pollution to identify area-specific trends across the NE US from 2000 to 2014. Aim 3 extends these small scale methods to a case study of Pennsylvania (PA) to investigate how the fracking industry has influenced $PM_{2.5}$ pollution variability across PA.

This research introduces an innovative approach for comparing air pollution maps between two time periods. We provide evidence that supports power plant emission controls as effective tools to reduce $PM_{2.5}$ pollution. This research contributes to the understanding of the extent of $PM_{2.5}$ pollution in the NE US by identifying the small scale variability in $PM_{2.5}$ trends and highlights the need for

improved $PM_{2.5}$ monitoring, particularly in areas at risk for air pollution from activities such as fracking.

Advisor:

Frank C. Curriero, PhD	Department of Epidemiology
	Department of Biostatistics

Thesis Committee:

Darryn W. Waugh, PhD (*Chair*)	Department of Earth and Planetary Sciences
	(Krieger School of Arts and Sciences)
Kirsten A. Koehler, PhD	Department of Environmental Health Sciences
Meghan F. Davis, PhD	Department of Environmental Health Sciences
Mary A. Fox, PhD	Department of Health Policy and Management
Ana M. Rule, PhD	Department of Environmental Health Sciences

Dedicated to my parents, with love and gratitude

Jumping from failure to failure

with undiminished enthusiasm

is the big secret to success.

- Savas Dimopoulos

ACKNOWLEDGEMENTS

I am greatly indebted to many, many people who contributed their time, knowledge, resources, understanding, and support during my tenure at the Johns Hopkins Bloomberg School of Public Health. First and foremost, I cannot thank my advisor, Dr. Frank Curriero, enough for his endless patience and expert guidance. The qualities that make him one of the finest professors at Johns Hopkins are amplified in his role as mentor; his passion for spatial analysis inspired me to pursue this path as a doctoral student, and his enthusiasm over the last five years motivated me to continue the pursuit even in the midst of academic and personal hardships. Frank is truly a mentor and a friend, and the completion of this dissertation document emanates directly from his dedication, work, and unwavering support.

I have had the privilege of receiving funding from the C. Sylvia and Eddie C. Brown Community Health Scholarship Program for the last five years. Eddie and Sylvia Brown provide extraordinary financial support to doctoral students pursing public health studies and allow the Brown Scholars unparalleled autonomy in their academic pursuits. They are actively engaged with their scholars, attending our seminars and inviting us to dine with them regularly. In addition to the support and encouragement the Browns provide, the Brown Scholarship establishes a tight-knit network of Brown Scholars under the tutelage of Dr. Robert (Bob) Blum with the assistance of Rachel Bass; the encouragement of Bob, Rachel, and my fellow Brown Scholars, particularly of Dr. Aracelis Torres and soon-to-be-Dr. Karina Christiansen,

has been a vital source of strength and fellowship that bolstered me through the pursuit of the doctoral degree.

The Environmental Health Sciences (EHS) department houses dedicated and wonderful faculty, staff, and students. Dr. Meghan Davis, Dr. Ana Rule, Dr. D'Ann Williams, and Christine Torrey were sources of expertise and kindness during these often very difficult years. I thank Dr. Kirsten Koehler, Dr. Peter Lees, Dr. Pat Breysse, Dr. Keeve Nachman, Courtney Mish, and Ruth Quinn for their guidance and their patience. EHS has also contributed funding for my doctoral training; I thank Peter particularly in his advocacy for this support. My fellow EHS students inspired me to work hard and also to have fun during my doctoral program. Dr. Sutyajeet (Sut) Soneja, Dr. Pam Dopart, Dr. Cissy Li, Dr. Patrick Baron, Dr. Jesse Berman, and Ben Davis deserve special thanks; I am so grateful for all of your help and for your friendship over the past five years.

Professors and colleagues across Johns Hopkins University and beyond provided training on the fundamental theories and practices underlying the research presented in this dissertation. I thank Dr. Thomas (Tom) Burke, Dr. Darryn Waugh, Dr. Marie Diener-West, Dr. Elizabeth Colantuoni, Dr. Mary Fox, Dr. Darcy Phelan-Emrick, Roger Messick, Rebecca Ruggles, and Tim Shields for their instruction and assistance. I am especially indebted to Tom for his training and guidance during my master's and doctoral programs at Johns Hopkins, and to Darryn, who helped me conceptualize this dissertation project and served as an active and engaged Chair of my research committee.

I was privileged to participate in the Christine Mirzayan Science and Technology Policy Graduate Fellowship Program at the National Academies of Sciences during my doctoral tenure. I thank Dr. Anne-Marie Mazza, Dr. James Reisa, Dr. Raymond Wassel, and Karolina Konarzewska for the unparalleled opportunity the Mirzayan Fellowship granted me, and I thank my fellow "Mirzayan 2015-ers" for their friendship and for their support during this final year of my doctoral studies.

And finally, to my family, and to my friends who have become my family: Mum, Dad, Truv, Nicole, Ben, Nan, Rachel, Darcy, Keri, Caitlin, and Jen: I could not have done this without you. I love you all.

TABLE OF CONTENTS

LIST OF TABLES

APPENDIX A

LIST OF FIGURES

CHAPTER 1: INTRODUCTION: SPECIFIC AIMS AND OVERVIEW OF PARTICULATE MATTER, FEDERAL AIR QUALITY REGULATIONS, AND SPATIAL ANALYSIS

Abbreviations

ARP	Acid Rain Program (Title IV-A of the 1990 Amendments to the CAA)
CAA	Clean Air Act
CAIR	Clean Air Interstate Rule
CSAPR	Cross-State Air Pollution Rule
EPA	United States Environmental Protection Agency
GIS	Geographic Information System
NAAQS	National Ambient Air Quality Standards
NE US	Northeastern United States
NOx	Nitrogen oxides
PM	Particulate matter
$PM_{2.5}$	Fine particulate matter, < 2.5 μm in aerodynamic diameter
PM_{10}	Coarse particulate matter, < 10 μm in aerodynamic diameter
SO_2	Sulfur dioxide
US	United States of America

1.1 Introduction

The health effects of particulate pollution exposure are well established and include increased risk of respiratory illness, aggravation of COPD, bronchitis, asthma, chest pain, and premature mortality (Dockery, Speizer et al. 1989, Dockery, Pope et al. 1993, Pope III, Thun et al. 1995, Ostro and Chestnut 1998, Laden, Neas et al. 2000, Peng, Bell et al. 2009). Particulate matter (PM) pollution consists of solid and liquid droplet particles suspended in air (United States Environmental Protection Agency 2016). People are exposed to PM by inhaling the particles, and PM of smaller particle sizes travel further and more efficiently through the body contributing to greater health impacts than larger sized PM (Laden, Neas et al. 2000). Particles less than 2.5 μm in aerodynamic diameter constitute fine particulate matter ($PM_{2.5}$).

Since 2000, national levels of ambient (outdoor) $PM_{2.5}$ concentrations have decreased (United States Environmental Protection Agency 2016). A recent study confirmed the northeastern United States (NE US) followed the national trend, with decreasing levels of $PM_{2.5}$ concentrations from 1999 – 2013 (Saunders and Waugh 2015). The research presented in this dissertation document builds upon the work of Saunders and Waugh (2015) to investigate how the variation of $PM_{2.5}$ pollution changes over space and time in the NE US. We examine the dynamic relationship between environmental determinants and $PM_{2.5}$ pollution. We assess the relative effects of federal regulations designed to decrease $PM_{2.5}$ pollution from 2000 - 2014. We apply an innovative approach to assess the small scale variability of $PM_{2.5}$ pollution to identify unique trends in $PM_{2.5}$ pollution levels across the NE US over

this time period. Finally, we extend the methods to a case study of Pennsylvania (PA), which experienced a fracking boom in the midst of the study period, to investigate how the fracking industry has influenced $PM_{2.5}$ pollution variability across PA.

1.2 Particulate Matter

Both stationary and mobile sources contribute to $PM_{2.5}$ pollution (Paciorek, Yanosky et al. 2009). The majority of $PM_{2.5}$ constituent particles result from combustion activities from motor vehicles and the burning of fossil fuels (Laden, Neas et al. 2000, Paciorek, Yanosky et al. 2009). $PM_{2.5}$ may also contain small particles from pulverized road dust, soil, and other grinding and crushing products from industry, agriculture, road systems and use, and other sources (Laden, Neas et al. 2000). In addition to these anthropogenic sources, natural sources contribute to $PM_{2.5}$ pollution; specifically, wildfires are a significant source of $PM_{2.5}$ in the US (Saunders and Waugh 2015).

Local daily concentrations of $PM_{2.5}$ are influenced strongly by meteorological variations, including temperature, relative humidity, precipitation, and circulation (Hand, Schichtel et al. 2012). Long range transportation of particulates also impacts local $PM_{2.5}$ ambient concentrations. Particulate pollution may travel into US airspace from Canada and Mexico, from off shore shipping activities, and even across oceans: Hand, Schichtel et al. (2012) notes high particulate events in Asia affect dust and other PM concentrations in the US (Hand, Schichtel et al. 2012). Particulate

emissions in the eastern US have been shown to influence the ambient trends across the entire country (Hand, Schichtel et al. 2012).

Particulate Matter and Human Health

Particulate pollution has long been associated with adverse health effects. The rapid expansion of industry including coal-fired industrial facilities and power stations, steel mills, coke ovens, foundries, and smelters resulted in numerous lethal smog events across industrialized countries throughout the 20th century (Bell and Davis 2001). Famously, the Great Smog of 1952 blanketed London in a thick layer of particulate-dense pollution for days, causing 12,000 excess deaths and hundreds of thousands more to fall ill (Bell and Davis 2001). The Great Smog brought public awareness to the health effects of air pollution and inspired regulatory actions to protect the public's health from the risks of air pollution (Bell, Davis et al. 2004).

The Harvard Six Cities Study provided prospective cohort data supporting the associations between particulate pollution and human health risks (Dockery, Pope et al. 1993). 8,111 adults were enrolled from six cities across the US and followed prospectively from 1974 through 1991. Concurrently, an air monitoring station centrally located in each community provided data on the ambient concentrations of PM and other air pollutants including sulfur dioxide, ozone, and sulfates. The city-specific mortality rates were found to be associated with average levels of air pollutions even after adjusting for risk factors including smoking status, education level, and body mass index. The study identified the significant association between fine particulate pollution ($PM_{2.5}$) and mortality due to lung cancer and cardiopulmonary disease.

Since the publications of the Harvard Six Cities Study, evidence has mounted linking $PM_{2.5}$ pollution with cardiovascular and respiratory effects and other causes of increased morbidity and mortality (Dockery, Speizer et al. 1989, Pope III, Thun et al. 1995, Schwartz, Dockery et al. 1996, Pope III and Dockery 2006, Miller, Siscovick et al. 2007, Peng, Bell et al. 2009). Dockery et al. (1989) identified elevated rates of respiratory illnesses among children living in high PM areas. Children with asthma are particularly susceptible to health impacts from $PM_{2.5}$ pollution (Delfino, Quintana et al. 2004). Elderly people exposed to particulate pollution are at risk for increased blood pressure (Delfino, Tjoa et al. 2010), and smokers are at risk for increased cardiovascular effects of exposure to $PM_{2.5}$ pollution (Pope, Burnett et al. 2004). Evidence suggests that the severity of health effects associated with $PM_{2.5}$ exposure may be source-dependent: mobile and combustion sources produce key constituents associated with health effects including respiratory and cardiovascular disease (Laden, Neas et al. 2000, Peng, Bell et al. 2009). Recent studies have linked $PM_{2.5}$ pollution with non-lung cancers, including cancers of the breast and upper digestive tract (Wong, Tsang et al. 2016), and with preterm births, low birth weight, and other adverse birth outcomes (Laurent, Hu et al. 2016, Pedersen, Gehring et al. 2016, Trasande, Malecha et al. 2016).

Pollution and Exposure Estimation

Epidemiology studies of the health effects associated with $PM_{2.5}$ exposure traditionally reply on measurements taken by ground-level air quality monitors (Kloog, Chudnovsky et al. 2014). This method may lead to exposure misclassification as the $PM_{2.5}$ concentration levels are extrapolated from the

stationary monitor to the ambulatory population of interest (Kloog, Chudnovsky et al. 2014, Berman, Breysse et al. 2015). Exposure estimates for persons living and working far from $PM_{2.5}$ monitors are even more prone to exposure classification errors. Statistical modeling, including land use regression and kriging, may reduce exposure misclassification by using information about the exposure area, such as the physical and built environment, and the distance from the nearest monitor to enhance predictive power (Kloog, Chudnovsky et al. 2014, Berman, Breysse et al. 2015). Kriging is discussed further below (see "Spatial Data and Statistical Methods: Methods of Interpolation"). Other methods of $PM_{2.5}$ exposure or pollution level estimation include satellite-based aerosol optical depth (AOD) and the use of surrogate visibility measurements such as visible range to estimate particulate concentrations (Paciorek, Yanosky et al. 2009, Kloog, Chudnovsky et al. 2014).

The United States Environmental Protection Agency (EPA) collects ambient $PM_{2.5}$ concentration data from ground-level air quality monitors as part of the nationwide Air Quality System (AQS). Convenience and knowledge of individual pollutants inform the placement of air monitors in the AQS network (Paciorek, Yanosky et al. 2009, Le, Breysse et al. 2014). Some monitors are placed in areas of high pollution but low population, with the intent to quantify pollution rather than exposure levels, while other monitors are strategically placed based on population density to best capture estimates of population exposure to air pollutants (Paciorek, Yanosky et al. 2009). Current and historical data from the AQS is available for public download and use from the EPA AQS website: http://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html.

1.3 Federal Regulations to Reduce Particulate Pollution

Under the authority of the Clean Air Act (CAA), the EPA enacts regulations to protect the public health and welfare from air pollution (United States Environmental Protection Agency 2015). EPA classifies both coarse PM (PM_{10}) and fine PM ($PM_{2.5}$) as *criteria air pollutants* since PM pollution is ubiquitous across the US and poses risks to the public's health and the environment (Dockery, Speizer et al. 1989, Pope III, Thun et al. 1995, Laden, Neas et al. 2000, Peng, Bell et al. 2009, Hasheminassab, Daher et al. 2014, United States Environmental Protection Agency 2016). EPA sets national air quality standards (NAAQS) and collects air samples via the AQS to assess attainment. NAAQS include limits to protect human health (*primary standards*) and to protect public welfare (*secondary standards*), which includes protections for visibility, animal and vegetation welfare, and infrastructure (United States Environmental Protection Agency 2016). The CAA requires individual states to design implementation plans and control measures so that the states attain NAAQS for criteria air pollutants including $PM_{2.5}$ (Hasheminassab, Daher et al. 2014).

EPA revises NAAQS for $PM_{2.5}$ and other criteria air pollutants as new data supports revisions of the standards and new technologies allow for additional air pollution controls (Table 1). The initial PM NAAQS, passed in 1971, set an annual geometric mean standard for total suspended particles (TSP) at 75 $\mu g/m^3$ for the primary NAAQS (United States Environmental Protection Agency 2016). The first standards specifically for $PM_{2.5}$ were enacted in 1997 and set the primary standard annual arithmetic mean at 50 $\mu g/m^3$ (Final Rule 62 FR 38652, Jul 18, 1997). The primary standard fell in 2006 to an annual mean of 15.0 $\mu g/m^3$ (Final rule 71 FR

61144, Oct 17, 2006) and again in 2012 to an annual mean of 12.0 μg/m^3 (Final rule 78 FR 3086, Jan 15, 2013). Attainment of standards considers annual means averaged over 3 years to assess compliance (United States Environmental Protection Agency 2016). In addition to the annual mean standards, primary standards include daily (24-hour) limits. The 2012 NAAQS $PM_{2.5}$ revisions set the daily limits at 35 μg/m^3 (Table 1). The daily NAAQS is calculated as a 3 year average of the 98th percentile concentration to assess compliance (United States Environmental Protection Agency 2016). Assessment of daily and annual NAAQS attainment is based on EPA's system of air quality monitors in the AQS.

Power Plant Regulations

To meet the limits of the $PM_{2.5}$ NAAQS, EPA passes regulations to reduce emissions that contribute to $PM_{2.5}$ air pollution (Figure 1). Power plants are the dominant source of sulfur dioxide (SO_2) and nitrogen oxides (NOx), which are precursor emissions of $PM_{2.5}$ (Hand, Schichtel et al. 2012, De Gouw, Parrish et al. 2014). The 1990 Amendments to the CAA included power plant emissions control regulations for SO_2 and NOx in Title IV – A, Acid Deposition Control (the Acid Rain Program, ARP) (United States Environmental Protection Agency 2015). In Title IV / ARP, EPA developed market-based approaches to meet emission reduction goals of SO_2 and NOx from power plants (U.S. Government Accountability Office 2005). SO_2 emission regulations under Title IV / ARP included a trading program, allowing utilities to meet the requirements through outright emission reductions or a combination of emission reductions and buying, selling, trading, or banking emissions to meet the annual allowances (U.S. Government Accountability Office

2005). Title IV / ARP also created continuous monitoring requirements for utilities to assess emissions. A 2005 report by the U.S. Government Accountability Office (GAO) named Title IV / ARP an "unqualified success" in reducing SO_2 and NOx emissions from power plants since 1990, a conclusion echoed by publications investigating ambient levels of these $PM_{2.5}$ precursors (U.S. Government Accountability Office 2005, Hand, Schichtel et al. 2012, De Gouw, Parrish et al. 2014).

Other notable federal regulations designed to reduce SO_2 and NOx emissions from power plants include the Clean Air Interstate Rule (CAIR) and the Cross-State Air Pollution Rule (CSAPR) (Figure 1). CAIR, enacted on March 10, 2005, required fossil fuel fired power plants to reduce SO_2 emissions beyond the requirements of Title IV / ARP over two deadlines (2010 and 2015), and also to reduce NOx emissions over deadlines in 2009 and 2015 (Indiana Department of Environmental Management 2015, United States Environmental Protection Agency 2016). CAIR employed an interstate cap and trade program modeled on Title IV / ARP to meet these emission reductions, and replaced the NOx Budget Trading Program (NBP) that had operated from 2003 - 2008 (Sotkiewicz and Holt 2005, United States Environmental Protection Agency 2016, United States Environmental Protection Agency 2016). In December 2008, the United States Court of Appeals for the D.C. Circuit determined that while the CAIR requirements were legal, the methods enumerated in the regulation were flawed, and directed EPA to issue a replacement rule (United States Environmental Protection Agency 2011, Indiana Department of Environmental Management 2015). Following this ruling, EPA enacted CSAPR on July 6, 2011 (United States Environmental Protection Agency 2016). CSAPR focused

on 25 states in the eastern US and required power plants in those states to reduce emissions of SO_2 and NOx, with the explicit objective of reducing pollution of $PM_{2.5}$ precursors that cross state boundaries and impact the air quality of neighboring states. Court rulings from the D.C. Circuit and the U.S. Supreme Court delayed the implementation of CSAPR, but on October 23, 2014, the D.C. Circuit granted EPA's request to lift the postponement of CSAPR implementation (Indiana Department of Environmental Management 2015, United States Environmental Protection Agency 2016). In 2015, Phase I of CSAPR, which reduced SO_2 emissions as well as both annual and ozone-season NOx emissions, took effect (United States Environmental Protection Agency 2016). In 2017, EPA plans to implement Phase II, which targets specific states for further SO_2 emissions (United States Environmental Protection Agency 2016).

Future reductions of $PM_{2.5}$ precursor emissions from power plants may come from the Mercury and Air Toxics Standards (MATS), finalized by EPA on December 16, 2011 (United States Environmental Protection Agency 2016). The goals of MATS include a 41% reduction in SO_2 emissions from power plants compared to the CSAPR limits. However, MATS has been challenged in court since its inception; on June 29, 2015, the U.S. Supreme Court ruled that the EPA must edit MATS to consider the cost of implementing its emission regulations (2015).

Mobile Source Regulations

EPA also passes regulations aimed at reducing $PM_{2.5}$ and precursor emissions from mobile sources, which are the other major contributor to $PM_{2.5}$ pollution in addition to power plants (Gillies and Gertler 2000, Greco, Wilson et al. 2007,

Hasheminassab, Daher et al. 2014). The combustion engines in cars, trucks, trains, airplanes, and other mobile sources result in both direct $PM_{2.5}$ emissions and indirect emissions of precursors including SO_2, NOx, and hydrocarbons (Greco, Wilson et al. 2007). EPA estimates that cars and trucks account for half of the emissions of NOx in urban areas (United States Environmental Protection Agency 2015).

To control $PM_{2.5}$ and other pollutants from mobile sources, EPA regulates both vehicle emissions and fuel quality. Title II of the 1990 CAA Amendments asserted tighter standards for car and truck tailpipe emissions beginning with automobiles in model year 1994 (Figure 1) (United States Environmental Protection Agency 2015). Title II also introduced requirements for "cleaner" gasoline, with lower volatility and sulfur content. On February 10, 2000, EPA finalized the Tier 2 Motor Vehicle Emissions Standards and Gasoline Sulfur Control Requirements (Tier 2 standards) which further constricted passenger vehicle emissions and gasoline quality to control pollution (Figure 1) (United States Environmental Protection Agency 2000). Under Tier 2 standards, cars, trucks, and SUV's were required to meet the stricter tailpipe emission standards of 0.07 grams per mile for nitrogen oxides beginning with model year 2004 passenger vehicles (United States Environmental Protection Agency 1999). Tier 2 standards tackled fuel quality by capping sulfur levels in gasoline. From 2004 – 2007, gasoline refiners and importers were required to meet the new standard of 30 ppm average sulfur content, with a maximum sulfur content not to exceed 80 ppm (United States Environmental Protection Agency 1999).

Tier 2 standards focused on passenger vehicles and unleaded gasoline, while the Heavy-Duty Engine and Vehicle Standards and Highway Diesel Fuel Sulfur Control Requirements (heavy-duty engine / diesel standards), adopted in 2000, asserted emissions limits for heavy-duty (non-passenger) vehicles and diesel gasoline standards (Figure 1) (United States Environmental Protection Agency 2016). Under the heavy-duty engine / diesel standards, trucks were required to include a diesel particulate filter as well as NOx exhaust-control technology, starting with model years 2007 – 2010, while diesel fuel sulfur content was capped at 15 ppm (Manufacturers of Emission Controls Association 2016). Future reductions on $PM_{2.5}$ precursor emissions from mobile sources are expected with the Tier 3 Vehicle Emissions and Fuel Standards Program (Tier 3 standards), which will further constrict tailpipe emissions allowances as well as unleaded gasoline sulfur content beginning in 2017 (United States Environmental Protection Agency 2016).

1.4 Spatial Data and Statistical Methods

Spatial Data

Data that are associated with locations are *spatial data*. Spatial data may be *point pattern data*, in which events are tagged with locations such as geographic coordinates (i.e. latitude and longitude) (Schabenberger and Gotway 2005, Bivand, Pebesma et al. 2008). Point pattern data can be aggregated across space and expressed as a count or rate associated with a boundary defined by location; such spatial data is called *areal data* (Bivand, Pebesma et al. 2008). Spatial data that could

in theory be measured at any location in a study area but are expressed by the finite measurements of the study are *geostatistical data* (Bivand, Pebesma et al. 2008).

The research described in this dissertation document utilizes all three types of spatial data. The outcome of interest, $PM_{2.5}$ concentration levels, is geostatistical: $PM_{2.5}$ pollution is ubiquitous, but the $PM_{2.5}$ dataset is geographically limited by the locations of the $PM_{2.5}$ monitors in the AQS network, and $PM_{2.5}$ concentration data points are tagged with the locations of the monitors. The power plant locations considered in the models explored in Chapters 2 and 3 are point pattern data, with the event (a power plant in the study area) tagged with event location information (latitude and longitude). The state- and county-level covariates considered in the models in Chapters 2 and 3 are examples of areal data, with counts or rates expressed over the areal boundaries. For example, the state-level traffic covariate is expressed in billions of vehicle miles traveled per state square mile, and traffic information is tagged with the geographical boundaries of the associated state. The county-level population covariate is expressed as thousands of people per county square mile, and the information is tagged with the geographic boundaries of the associated county.

Methods of Interpolation

Statistical estimations of spatial data like $PM_{2.5}$ pollution levels should consider spatial characteristics. It is expected that $PM_{2.5}$ data exhibit *spatial autocorrelation*: $PM_{2.5}$ concentration levels are more similar for locations close together compared to locations further apart (Schabenberger and Gotway 2005). Statistical models aimed at predicting $PM_{2.5}$ pollution levels at unmonitored

locations should beware the assumption of *homoscedasticity* (constant variance). Model residuals must be analyzed for *residual spatial variation* to test this assumption.

The construction of residual semivariograms provides a method to investigate residual spatial variation. A horizontal semivariogram indicates spatial independence (no residual spatial variation) (Cressie 1993, Verdú and García‐Fayos 1998, Mannshardt-Shamseldin, Smith et al. 2010). As previously noted, $PM_{2.5}$ data displays spatial autocorrelation. If the addition of covariates into a regression model accounts for the spatial variation of the $PM_{2.5}$ data, then the semivariogram of the model residuals will be flat, indicating that the error term of the model is normally distributed about 0 and displays constant variance σ^2:

$$e \sim N(0,\sigma^2)$$

The spatial analysis of the multilevel model considered in Chapter 2 of this dissertation document demonstrates how the addition of regression covariates accounts for the spatial variation of $PM_{2.5}$ in the NE US (Chapter 2, Figure 3).

If covariates are not considered or if the addition of regression covariates fails to account for the spatial variation of the $PM_{2.5}$ data, then the semivariogram of the model residuals will indicate the error term of the model varies spatially:

$$e \sim N(0,\sigma^2\mathbf{R}) \quad (1)$$

where the variance σ^2 is multiplied by a distance matrix **R** to account for the spatial structure of the residuals wherein the spatial correlation of the residuals decreases as a function of distance between locations (Berman, Breysse et al. 2015). The spatial analysis of the null model (no regression covariates) considered in Chapter 2 of this dissertation document demonstrates a semivariogram showing residual spatial variation (Chapter 2, Figure 2).

Kriging is a regression-based spatial interpolation methods that provides the best linear unbiased prediction (BLUP) while allowing the error term of the model to vary spatially (formula 1) (Schabenberger and Gotway 2005). *Ordinary kriging* assumes a constant mean across locations (Schabenberger and Gotway 2005). Thus, the ordinary kriging formula includes only the intercept of the regression model:

$$PM_i = \beta_0 + e_i$$

where PM_i is the $PM_{2.5}$ concentration at monitor *i*, β_0 is the constant mean, and e_i is the residual error term, which varies spatially (formula 1). *Universal kriging* extends the ordinary kriging model to include covariates:

$$PM_i = \beta_0 + \beta_1 X_1 \ldots + \beta_n X_n + e_i$$

where, again, the residual error term varies spatially (formula 1). Kriging yields the best linear unbiased predictions and provides a measure of prediction uncertainty in the minimized mean square prediction errors (Schabenberger and Gotway 2005,

Berman, Breysse et al. 2015). Chapters 3 and 4 of this dissertation document utilize kriging to complete air pollution maps of $PM_{2.5}$.

Software

Geographic information systems (GIS) are specialized software for the collection, storage, transformation, and display of spatial data (Bivand, Pebesma et al. 2008). The work presented in this dissertation utilizes QGIS (version 2.10.1-Pisa) to store and map spatial data. Statistical analyses are completed using the R statistical software (version 3.2.3), employing packages including geoR and sp for spatial analysis, lme4, nlme, lmerTest, car, and MASS for regression and mixed effects modeling, and lattice for data visualization.

1.5 Dissertation Objective and Specific Aims

The objective of this dissertation is to characterize the variation in $PM_{2.5}$ over space and time in the northeastern United States (NE US) from 2000 to 2014. To accomplish this, we propose three specific research aims:

Specific Aim 1: To characterize the spatial-temporal variation in $PM_{2.5}$ in the NE US from 2000-2014 as a function of environmental determinants.

We hypothesize that significant environmental determinants that influence $PM_{2.5}$ concentrations include monitor proximity to a power plant, net energy generation,

and traffic density, because power plants and mobile sources are the primary sources of $PM_{2.5}$ and precursor emissions.

Specific Aim 2: To investigate large scale trends and small scale spatial variation in $PM_{2.5}$ pollution and the efficacy of federal emissions regulations in reducing $PM_{2.5}$ pollution in the NE US.

We hypothesize that the national trend of decreasing $PM_{2.5}$ concentrations will be affirmed in the NE US, but that our investigation will identify smaller scale variations, including regions within the NE US that did not experience a significant decrease in $PM_{2.5}$ concentrations from 2000 to 2014 despite the large scale trend.

Specific Aim 3: To explore whether the establishment of the fracking industry in Pennsylvania (PA) impacted the small scale spatial variability in $PM_{2.5}$ pollution within the state from 2004 to 2014.

We hypothesize that different regions within PA experienced different trends in $PM_{2.5}$ pollution during this time period, and that the presence of fracking industry impacts these small scale trends.

1.6 Dissertation Document Organization

Throughout this dissertation document, the term "airscape" is used to encompass the ambient air quality features of an area. Other terminologies are

defined throughout the document. Common abbreviations are listed in the beginning of each chapter.

The specific research aims enumerated above align with distinct chapters in this dissertation document. Chapter 2 presents research Aim 1, and Appendix A contains supplemental materials from Chapter 2. Chapter 3 encompasses research Aim 2, with supplemental materials in Appendix B. Chapter 4 illustrates research Aim 3, and Appendix C contains supplemental materials from this chapter. Chapter 5 concludes the dissertation with a review of the research aims and findings as well as the research strengths and limitations. It concludes with a discussion of the public health and research contributions and the consideration of future research directions.

1.7 Tables and Figures

Year	Primary Standard (annual)	Primary Standard (24 hour)
1971	TSP = 75 µg/m3	260 µg/m3
1987	PM10 = 50 µg/m3	150 µg/m3
1997	15.0 µg/m3	65 µg/m3
2006	15.0 µg/m3	35 µg/m3
2012	12.0 µg/m3	35 µg/m3

Table 1. A history of NAAQS for $PM_{2.5}$

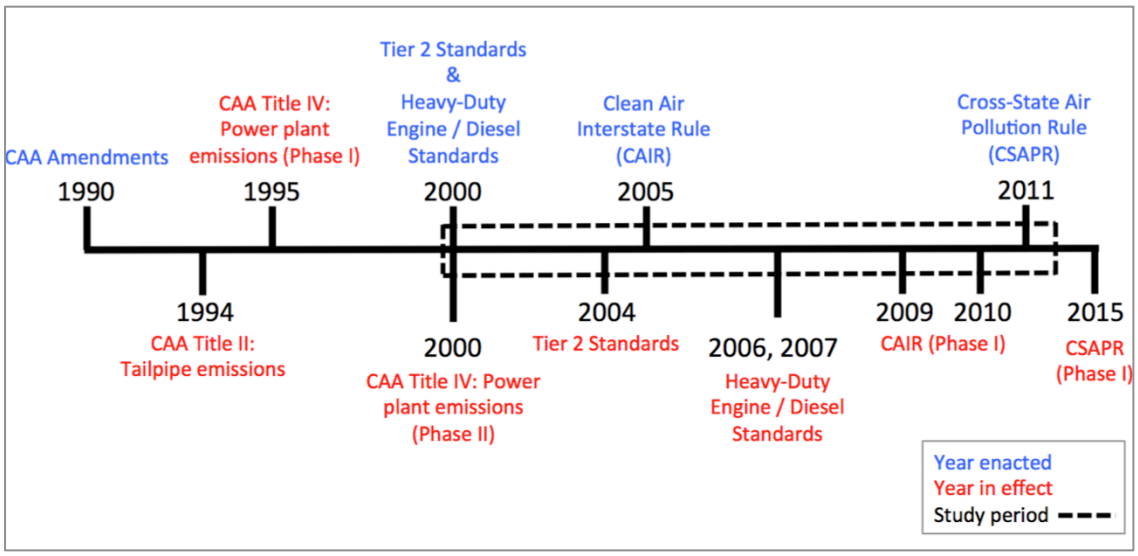

Figure 1. Timeline of select federal regulations aimed at curbing $PM_{2.5}$ and $PM_{2.5}$ precursor emissions. The study period indicates 2000 – 2014, the years investigated in this dissertation.

1.8 References

(2015). Michigan et al. V. Environmental Protection Agency et al., United States Supreme Court.

Bell, M. L. and D. L. Davis (2001). "Reassessment of the lethal London fog of 1952: novel indicators of acute and chronic consequences of acute exposure to air pollution." Environmental Health Perspectives **109**(Suppl 3): 389.

Bell, M. L., et al. (2004). "A retrospective assessment of mortality from the London smog episode of 1952: the role of influenza and pollution." Environmental Health Perspectives **112**(1): 6.

Berman, J. D., et al. (2015). "Evaluating methods for spatial mapping: Applications for estimating ozone concentrations across the contiguous United States." Environmental Technology & Innovation **3**: 1-10.

Bivand, R. S., et al. (2008). "Applied spatial data analysis with R. Springer."

Cressie, N. (1993). "Statistics for spatial data: Wiley series in probability and statistics." Wiley-Interscience New York **15**: 16.

De Gouw, J., et al. (2014). "Reduced emissions of CO2, NOx, and SO2 from US power plants owing to switch from coal to natural gas with combined cycle technology." Earth's Future **2**(2): 75-82.

Delfino, R. J., et al. (2004). "Association of FEV1 in asthmatic children with personal and microenvironmental exposure to airborne particulate matter." Environmental Health Perspectives **112**(8): 932.

Delfino, R. J., et al. (2010). "Traffic-related air pollution and blood pressure in elderly subjects with coronary artery disease." Epidemiology (Cambridge, Mass.) **21**(3).

Dockery, D. W., et al. (1993). "An association between air pollution and mortality in six US cities." New England journal of medicine **329**(24): 1753-1759.

Dockery, D. W., et al. (1989). "Effects of inhalable particles on respiratory health of children." American Review of Respiratory Disease **139**(3): 587-594.

Gillies, J. A. and A. W. Gertler (2000). "Comparison and evaluation of chemically speciated mobile source PM2. 5 particulate matter profiles." Journal of the Air & Waste Management Association **50**(8): 1459-1480.

Greco, S. L., et al. (2007). "Spatial patterns of mobile source particulate matter emissions-to-exposure relationships across the United States." Atmospheric Environment **41**(5): 1011-1025.

Hand, J., et al. (2012). "Particulate sulfate ion concentration and SO 2 emission trends in the United States from the early 1990s through 2010." Atmospheric Chemistry and Physics **12**(21): 10353-10365.

Hasheminassab, S., et al. (2014). "Long-term source apportionment of ambient fine particulate matter (PM 2.5) in the Los Angeles Basin: A focus on emissions reduction from vehicular sources." Environmental Pollution **193**: 54-64.

Indiana Department of Environmental Management (2015). "CAIR, Cross-State Air Pollution Rule (CSAPR) and Transport Rule Timeline." Retrieved 3/13/2016, from http://www.in.gov/idem/airquality/2557.htm.

Kloog, I., et al. (2014). "A new hybrid spatio-temporal model for estimating daily multi-year PM 2.5 concentrations across northeastern USA using high resolution aerosol optical depth data." Atmospheric Environment **95**: 581-590.

Laden, F., et al. (2000). "Association of fine particulate matter from different sources with daily mortality in six US cities." Environmental Health Perspectives **108**(10): 941.

Laurent, O., et al. (2016). "A Statewide Nested Case-Control Study of Preterm Birth and Air Pollution by Source and Composition: California, 2001–2008." Environ Health Perspect.

Le, G. E., et al. (2014). "Canadian forest fires and the effects of long-range transboundary air pollution on hospitalizations among the elderly." ISPRS International Journal of Geo-Information **3**(2): 713-731.

Mannshardt-Shamseldin, E. C., et al. (2010). "Downscaling extremes: A comparison of extreme value distributions in point-source and gridded precipitation data." The Annals of Applied Statistics: 484-502.

Manufacturers of Emission Controls Association (2016). "U.S. EPA 2007/2010 Heavy-Duty Engine and Vehicle Standards and Highway Diesel Fuel Sulfur Control Requirements." Retrieved 4/12/2016, from http://www.meca.org/regulation/us-epa-20072010-heavyduty-engine-and-vehicle-standards-and-highway-diesel-fuel-sulfur-control-requirements.

Miller, K. A., et al. (2007). "Long-term exposure to air pollution and incidence of cardiovascular events in women." New England journal of medicine **356**(5): 447-458.

Ostro, B. and L. Chestnut (1998). "Assessing the health benefits of reducing particulate matter air pollution in the United States." Environmental Research **76**(2): 94-106.

Paciorek, C. J., et al. (2009). "Practical large-scale spatio-temporal modeling of particulate matter concentrations." Annals of Applied Statistics **3**(1): 370 - 397.

Pedersen, M., et al. (2016). "Elemental constituents of particulate matter and newborn's size in eight European cohorts." Environmental Health Perspectives **124**(1): 141.

Peng, R. D., et al. (2009). "Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution." Environmental Health Perspectives **117**(6): 957.

Pope, C. A., et al. (2004). "Cardiovascular mortality and long-term exposure to particulate air pollution epidemiological evidence of general pathophysiological pathways of disease." Circulation **109**(1): 71-77.

Pope III, C. A. and D. W. Dockery (2006). "Health effects of fine particulate air pollution: lines that connect." Journal of the Air & Waste Management Association **56**(6): 709-742.

Pope III, C. A., et al. (1995). "Particulate air pollution as a predictor of mortality in a prospective study of US adults." American journal of respiratory and critical care medicine **151**(3_pt_1): 669-674.

Saunders, R. O. and D. W. Waugh (2015). "Variability and potential sources of summer PM 2.5 in the Northeastern United States." Atmospheric Environment **117**: 259-270.

Schabenberger, O. and C. A. Gotway (2005). Statistical methods for spatial data analysis, CRC press.

Schwartz, J., et al. (1996). "Is daily mortality associated specifically with fine particles?" Journal of the Air & Waste Management Association **46**(10): 927-939.

Sotkiewicz, P. M. and L. Holt (2005). "Public utility commission regulation and cost-effectiveness of Title IV: lessons for CAIR." The Electricity Journal **18**(8): 68-80.

Trasande, L., et al. (2016). "Particulate Matter Exposure and Preterm Birth: Estimates of US Attributable Burden and Economic Costs." Environ Health Perspect.

U.S. Government Accountability Office (2005). Clean Air Act: EPA Has Completed Most of the Actions Required by the 1990 Amendments, but Many Were Completed Late. http://www.gao.gov.

United States Environmental Protection Agency (1999, December). "Regulatory Announcement: EPA's Program for Cleaner Vehicles and Cleaner Gasoline." Retrieved 4/12/2016, from https://www3.epa.gov/tier2/documents/f99051.pdf.

United States Environmental Protection Agency (2000). Tier 2 Motor Vehicle Emissions Standards and Gasoline Sulfur Control Requirements EPA. Federal Register. **40 CFR Parts 80, 85, and 86**.

United States Environmental Protection Agency (2011, July 18). "The Cross-State Air Pollution Rule: Reducing the Interstate Transport of Fine Particulate Matter and Ozone (Fact Sheet)." Retrieved 4/3/2016, from https://www3.epa.gov/crossstaterule/pdfs/CSAPRFactsheet.pdf.

United States Environmental Protection Agency (2015, October 27). "1990 Clean Air Act Amendment Summary." Retrieved 3/4/2016, from https://www.epa.gov/clean-air-act-overview/1990-clean-air-act-amendment-summary.

United States Environmental Protection Agency (2015, October 27). "1990 Clean Air Act Amendment Summary: Title II." Retrieved 3/13/2016, from https://www.epa.gov/clean-air-act-overview/1990-clean-air-act-amendment-summary-title-ii.

United States Environmental Protection Agency (2015, October 27). "Clean Air Act Overview." Retrieved 3/31/16, from https://www.epa.gov/clean-air-act-overview.

United States Environmental Protection Agency (2016, February 21). "Clean Air Interstate Rule (CAIR)." Retrieved 3/13/2016, from https://archive.epa.gov/airmarkets/programs/cair/web/html/index.html.

United States Environmental Protection Agency (2016, February 29). "Cross-State Air Pollution Rule (CSAPR)." Retrieved 2/28/2016, from https://www3.epa.gov/crossstaterule/.

United States Environmental Protection Agency (2016, February 22). "Cross-State Air Pollution Rule (CSAPR) Basic Information." Retrieved 4/12/2016, from https://www3.epa.gov/crossstaterule/basic.html.

United States Environmental Protection Agency (2016, February 22). "Heavy-Duty Highway Diesel Program." Retrieved 4/12/2016, from https://www3.epa.gov/otaq/highway-diesel/regs.htm.

United States Environmental Protection Agency (2016, February 23). "Mercury and Air Toxics Standards (MATS): Cleaner Power Plants." Retrieved 4/12/2016, from https://www3.epa.gov/mats/powerplants.html.

United States Environmental Protection Agency (2016, February 23). "National Trends in Particulate Matter Levels." Retrieved 3/18/2016, from https://www3.epa.gov/airtrends/pm.html.

United States Environmental Protection Agency (2016, February 25). "NOx Budget Trading Program." Retrieved 4/12/2016, from https://www.epa.gov/airmarkets/nox-budget-trading-program.

United States Environmental Protection Agency (2016, February 23). "Particulate Matter (PM)." Retrieved 4/2/2016, from https://www3.epa.gov/airquality/particlepollution/.

United States Environmental Protection Agency (2016, April 6). "Reviewing National Ambient Air Quality Standards – Scientific and Technical Information." Retrieved 4/7/2016, from https://www3.epa.gov/ttn/naaqs/.

United States Environmental Protection Agency (2016, April 12). "Tier 3 Vehicle Emission and Fuel Standards Program." Retrieved 4/12/2016, from https://www3.epa.gov/otaq/tier3.htm.

Verdú, M. and P. García‐Fayos (1998). "Old‐field colonization by Daphne gnidium: seedling distribution and spatial dependence at different scales." Journal of Vegetation Science **9**(5): 713-718.

Wong, C. M., et al. (2016). "Cancer Mortality Risks from Long-term Exposure to Ambient Fine Particle." Cancer Epidemiology Biomarkers & Prevention.

CHAPTER 2: AIM I: THE CHARACTERIZATION OF $PM_{2.5}$ IN THE NORTHEASTERN UNITED STATES AS A FUNCTION OF ENVIRONMENTAL DETERMINANTS

Abstract

Research aim 1 employed non-spatial and spatial statistical modeling techniques to assess $PM_{2.5}$ pollution in the northeastern United States (NE US) to identify significant environmental determinants associated with $PM_{2.5}$ pollution. Previously identified environmental covariates at different spatial aggregations were considered, including monitor and power plant locations at the point-level, population, toxic releases, and elevation at the county level, and energy generation and annual traffic at the state level of spatial aggregation. The temporal covariates of season and year were also explored. We undertook a deliberate, step-wise approach to build our final model, comparing model performance and significance of covariates as well as investigating the model residual spatial dependence, to arrive at a statistically sound working model for $PM_{2.5}$ in the NE US. The construction of the best model is a critical first step in the analysis of $PM_{2.5}$ pollution. The work of this research aim determines the final model that we employ in subsequent research aims of this dissertation.

Abbreviations

AQS	Air quality system
CAA	Clean Air Act
EIA	United States Energy Information Administration
EPA	United States Environmental Protection Agency
GIS	Geographic information system
MWH	Megawatt hour
NE US	Northeast United States
NOx	Nitrogen oxides
PM	Particulate matter
$PM_{2.5}$	Fine particulate matter, < 2.5 μm in aerodynamic diameter
PM_{10}	Coarse particulate matter, < 10 μm in aerodynamic diameter
SO_2	Sulfur dioxide
TRI	Toxics release inventory
VMT	Vehicle miles traveled

2.1 Introduction

Electricity generation and mobile sources account for the majority of $PM_{2.5}$ pollution in the United States (US) (Laden, Neas et al. 2000, Greco, Wilson et al. 2007, Levy, Baxter et al. 2009). Power plants are the dominant source of sulfur dioxide (SO_2) and nitrogen oxides (NOx), which are precursor emissions of $PM_{2.5}$ (Hand, Schichtel et al. 2012, De Gouw, Parrish et al. 2014). Fossil fuel burning power plants emit ~69% of all SO_2 emissions, and SO_2 oxidizes into sulfate (SO_4^{2-}), the primary constituent of $PM_{2.5}$ in the northeastern US (NE US) (Marufu, Taubman et al. 2004, Paciorek, Yanosky et al. 2009, Hand, Schichtel et al. 2012). Hand, Schichtel et al. (2012) found that the NE US reported the highest annual SO_2 emissions from power plants in the country from 2000 – 2010.

SO_2 oxidizes slowly as it travels through the air from the power plant source, creating $PM_{2.5}$ as a secondary emission away from the point source (Paciorek, Yanosky et al. 2009). This behavior influences the long range transportation and the geographic reach of $PM_{2.5}$ and $PM_{2.5}$ precursors (Paciorek, Yanosky et al. 2009). Changes in power plant activities can radically influence the particulate airscape; for example, a state-wide decrease in SO_2 emissions in Washington state in 2002 can be traced to the impact of a single power plant that added SO_2 scrubbers and increased reliance on the lower-SO_2 emitting natural gas-fired units (Hand, Schichtel et al. 2012). While there has been an overall decreasing trend in power plant emissions of $PM_{2.5}$ precursors including SO_2 and NOx since the mid-1990's, there was an increase in emissions in the mid-2000's, while in 2008 - 2010, both emissions dropped in all areas of the United States except the southeast (Hand, Schichtel et al. 2012, De Gouw,

Parrish et al. 2014, Saunders and Waugh 2015). The economic recession has been identified as a potential cause of the emissions decrease in 2008 - 2010 (De Gouw, Parrish et al. 2014, Saunders and Waugh 2015). In addition to this precipitous drop, the decreasing trend in SO_2 and NOx emissions from power plants since 1995 has been due in part to the trend of the increasing use of natural gas in place of the traditional coal power (De Gouw, Parrish et al. 2014). Natural gas power plants emit less emissions than coal-fired plants due to the low sulfur content of natural gas and the efficient emission control technologies in a modern natural gas power plant (De Gouw, Parrish et al. 2014).

National policies have impacted the particulate pollution from power plant sources. Title IV of the 1990 Clean Air Act (CAA) Amendments required utilities to reduce SO_2 emissions by the year 2000, and the mandates resulted in a 60% decrease in total SO_2 emissions in the US from 1990 to 2010 (Hand, Schichtel et al. 2012, United States Environmental Protection Agency 2015). The Clean Air Interstate Rule (CAIR), signed into law on March 10, 2005, required fossil fuel fired power plants to further reduce their SO_2 emissions over two deadlines (2010 and 2015), and also to reduce NOx emissions (United States Environmental Protection Agency 2016). However, CAIR was replaced by the Cross-State Air Pollution Rule (CSAPR) on July 6, 2011; thus, the second CAIR deadline was replaced by the deadlines of CSAPR, which also required reductions in SO_2 and NOx emissions (Indiana Department of Environmental Management 2015, United States Environmental Protection Agency 2016).

In addition to power plants, mobile sources contribute significantly to $PM_{2.5}$ pollution (Gillies and Gertler 2000, Greco, Wilson et al. 2007, Hasheminassab, Daher et al. 2014). The combustion engines in cars, trucks, trains, airplanes, and other mobile sources result in both direct $PM_{2.5}$ emissions and indirect emissions of precursors including SO_2, NOx, and hydrocarbons (Greco, Wilson et al. 2007).

Non-power plant industries influence $PM_{2.5}$ concentrations through direct (primary emissions of $PM_{2.5}$) or indirect emissions (emissions of $PM_{2.5}$ precursors including SO_2) (Paciorek, Yanosky et al. 2009). Industrial processes that result in PM emissions include combustion and mechanical grinding and crushing (Laden, Neas et al. 2000, Paciorek, Yanosky et al. 2009). SO_2 emission sources include industrial, commercial, and institutional sources like heaters and boilers, chemical processes including chemical production, and petroleum refining (Hand, Schichtel et al. 2012).

The impact of emissions on ambient $PM_{2.5}$ concentrations is modified by season. Seasonality trends in ambient $PM_{2.5}$ concentrations differ across regions of the US but generally follow a bimodal distribution with peaks in the summer and winter seasons (Hand, Schichtel et al. 2012). In the NE US, $PM_{2.5}$ peaks in the summer are driven by wildfires and UV-driven photochemistry including solar insulation and high humidity that influences biogenic emissions (Hand, Schichtel et al. 2012, Kim, Jacob et al. 2015, Saunders and Waugh 2015). Saunders and Waugh (2015) identified a decreasing trend in both the magnitude and the variability of the summer $PM_{2.5}$ concentrations in the NE from 1999 – 2013.

Since 2000, ambient $PM_{2.5}$ concentrations have decreased across the US (United States Environmental Protection Agency 2016). Total SO_2 emissions have

also fallen, from 31 million tons in 1970 to 8 million tons in 2010 (Hand, Schichtel et al. 2012). Hand et al. (2012) identified reductions in power plant emissions as the primary driver of this decrease in total SO_2 emissions (Hand, Schichtel et al. 2012).

To investigate the influence of these factors on ambient $PM_{2.5}$ concentrations in the NE US from 2000 - 2014, we built a multilevel model that acknowledges the influence of space and time on $PM_{2.5}$ through the inclusion of space and time determinants. We hypothesize that significant environmental determinants influencing $PM_{2.5}$ concentrations include monitor proximity to a power plant, net energy generation, and traffic density, because power plants and mobile sources are the primary sources of $PM_{2.5}$ and precursor emissions. We anticipate that $PM_{2.5}$ monitor values exhibit *spatial autocorrelation*, meaning that $PM_{2.5}$ monitor values are more similar for monitors located close together compared to monitors located further apart (Schabenberger and Gotway 2005). Thus, we analyze our final model to determine if the inclusion of the model covariates accounted for the spatial autocorrelation phenomena. Failure to acknowledge the spatial component of the data ignores the processes underlying the outcome, and inferences based on a model that fails to account for spatial variation where it exists may underestimate standard errors (Cressie 1993, Schabenberger and Gotway 2005, Bivand, Pebesma et al. 2008, Berman, Breysse et al. 2015).

2.2 Methods

Study Area

We defined the NE US as extending from Virginia north through Maine, encompassing the following 14 states: Connecticut, Delaware, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, Virginia, Washington D.C., and West Virginia (Saunders and Waugh 2015) (Figure 1). The study area is 263,287.81 square miles (mi^2), with an average state size of 18,806.27 mi^2. Washington, D.C., is the smallest state in the study area (68.34 mi^2), and New York is the largest (54554.98 mi^2) (United States Census Bureau 2010). Our study area encompasses urban, suburban, industrial, and rural areas, with varied elevation.

The 14 states in the study area contain a total of 434 counties and county-equivalent areas as defined by the US Census Bureau (United States Census Bureau 2010). States average 31 counties, with Washington D.C. containing a single county and Virginia consisting of 134 counties and independent county-equivalent cities. The average size of the 434 counties in the study area is 606.65 mi^2 (range 2.00 - 6827.57 mi^2).

Data

$PM_{2.5}$

We downloaded daily summary $PM_{2.5}$ concentrations (Code 88101) for each year (2000 – 2014) from the EPA Air Quality System (AQS) website (United States

Environmental Protection Agency 2015). The AQS network of air monitors extends across the United States.

The AQS summary data includes the daily arithmetic mean $PM_{2.5}$ concentration for each monitor in the AQS system. The EPA defines the daily arithmetic mean as "the measure of central tendency obtained from the sum of the observed pollutant data values or National Ambient Air Quality Standards (NAAQS) averages in the daily data set divided by the number of values that comprise the sum for the daily data set. For criteria pollutants, the sum of values only adds the values with the appropriate flagging and concurrence for the exceptional data type"(United States Environmental Protection Agency 2011). Monitoring stations report either 24 hour average or hourly average PM concentrations (Paciorek, Yanosky et al. 2009).

For our analysis, we considered the daily $PM_{2.5}$ data for all monitors located in our study area, the NE US. For each monitor in our study, we averaged the daily data into a monthly average $PM_{2.5}$ value for that monitor.

Some monitors had missing or null values reported for the daily average. These missing observations may be due to equipment failures, maintenance, or the retirement of the monitoring site (Paciorek, Yanosky et al. 2009). Missing or null values were not considered in our analysis, and we assume that locations within our study area that do not have a $PM_{2.5}$ monitor in the AQS system (unsampled locations) are "missing at random", with no pattern to the missing pollution data. Paciorek, Yanosky et al. (2009) found missing observations to be missing completely

at random (MCAR), with no evidence of a pattern to the missing observations in a study using PM data from the AQS.

Power plants

Power plant locations were supplied by the EPA's Air Markets Program Data (AMPD) online system, accessible at ampd.epa.gov/ampd. The Acid Rain Program requires power plants over 25 megawatts to report to the EPA, which releases information to the public including power plant locations via the AMPD online system (Miller and Van Atten 2004). We downloaded the location information including latitude and longitude for all power plants in the US for all study years available (2001 – 2014). QGIS was used to build various buffer sizes around the power plants (5, 10, 50, and 100 km).

We tested the applicability of these different buffer sizes by running our final model on a subset of our data four times, with a different buffer size in each trial. For each trial model, we compared Akaike information criterion (AIC) and analysis of variance (ANOVA) p-values to determine the power plant buffer with the best model fit. AIC is a method of model selection that recognizes the tradeoff between bias and parsimony in statistical models: too few variables may bias the model inferences, while too many sacrifice model parsimony (Burnham and Anderson 2004).

We visually inspected all power plant study years as point values on a map in GIS. There were very slight yearly differences in power plant locations, and power plant location data was not available for study year 2000. Therefore, three representative years were selected for the analysis: power plant locations in 2001

were used to determine power plant proximity to PM monitor values for study years 2000 – 2003; power plant locations in 2006 were used for PM monitor values for 2004 – 2008; and power plant locations in 2011 were used for PM monitor values for 2009 – 2014.

Toxics release inventory

To capture a measurement of non-specific industrial emissions by county, the total releases of all chemicals in all industries was downloaded from the EPA's toxics release inventory (TRI) reports website, accessible at https://iaspub.epa.gov/triexplorer/tri_release.chemical. TRI compiles data submitted from facilities in different industrial sectors and allows public access of the data through the website (United States Environmental Protection Agency 2016). Under the Emergency Planning and Community Right-to-Know Act (EPCRA, 1986), facilities are required to report to EPA annual releases of toxic chemicals into the environment, including air, water, and underground emissions. We downloaded the total releases (pounds) for all chemicals and all industries for all facilities in our study area by county, from 2000 – 2014. We then generated a quintile rank for each county for each year, which compiled all releases by all facilities within a county for a given year and ranked the results, so that a county with relatively low total toxic releases ranks in the first quintile while a county with relatively high total toxic releases ranks in the fifth quintile.

Population

Previous studies have considered population density as a covariate in PM modeling (Greco, Wilson et al. 2007, Hart, Yanosky et al. 2009). A wide range of human activities, from home energy use to commuting, emits $PM_{2.5}$ and/or its precursors, and the population variable acts as a surrogate for these sources of $PM_{2.5}$.

We downloaded population estimates for each county in our study area from the U.S. Census Bureau's American Fact Finder website, accessible at http://factfinder.census.gov. The population data was generated from the 2000 U.S. Census (identification code DP-1) and from the American Community Surveys for 2005 – 2014 (identification code B01003). Data were not available for study years 2001 - 2004; census data from 2000 was used for study years 2001 – 2003, while census data from 2005 was used for study years 2004. We scaled the population covariate by county area to thousand people per county square mile:

$$\text{Population} = (\text{Population}/1000)/\ \text{County square miles} \quad (1)$$

Elevation

The U.S. Geological Survey (USGS) provides point-level elevation measurements across the United States (U.S. Geological Survey 2016). We used these point-level measurements to calculate a county average elevation, using all available data points in each county within our study area (meters).

Energy

We included the covariate of net energy generation by state to account for state-level emissions from the energy sector. A previous study considered this covariate in general additive models of PM_{10} across the United States (Hart, Yanosky et al. 2009). State-level monthly net energy generation data (thousand megawatt hours (MWH)) was retrieved from the US Energy Information Administration (EIA) for all fuel types and all energy sectors for all study years available (2001 – 2014) (U.S. Energy Information Administration 2015). EIA defines net generation as "the amount of gross generation less the electrical energy consumed at the generating station(s) for station service or auxiliaries. Electricity required for pumping at pumped-storage plants is regarded as electricity for station service and is deducted from gross generation" (U.S. Energy Information Administration). We scaled the energy use covariate to hundred thousand MWH per state square mile, calculated as:

$$\text{Energy} = (100 * \text{energy in thousand MWH}) / \text{State square miles} \quad (2)$$

Net energy generation was not available for year 2000; after investigating trends in energy use over time, we used the 2001 energy generation data as a surrogate for 2000 (Appendix A).

Traffic

To account for the effect of traffic-related $PM_{2.5}$ pollution in our model, we considered the annual vehicle miles traveled (VMT) per state as reported by the

Federal Highway Administration (FHWA) Office of Highway Policy Information. VMT statistics are gathered by the states and federal government for use in program planning and evaluation (Puentes and Tomer 2008). We utilized this dataset rather than estimating traffic densities (i.e. by dividing length of road by state or county areas) for a few reasons: first, the FHWA dataset is single-source and available for our entire study area from 2000 – 2011, and second, the FHWA dataset is a reflection of actual traffic, rather than using the presence of a road as a surrogate for traffic. Studies that use a surrogate measure for traffic, such as the presence of a road or distance to a road, are susceptible to misclassification bias, since the presence of a road does not necessarily correlate to road use (traffic). Similarly, characterizing the "road exposure" of a monitor as the distance from the monitor to a road is also sensitive to the assumption of traffic on the road (Hart, Yanosky et al. 2009).

We downloaded state-level VMT data (million VMT) from Table VM-2, located in the Roadway Extent, Characteristics and Performance section of the annual Highway Statistics Series, for 2000 - 2011 (Federal Highway Administration 2015). Data was not available for study years 2012 – 2014; FHWA VMT data from 2011 was used for those years. We scaled VMT into billion VMT per state square mile:

$$\text{Traffic} = (\text{Million VMT} / 1000) / \text{State square mile} \quad (3)$$

Season

We defined seasons as follows: winter spanned the months of December, January, and February; spring entailed March, April, and May; summer covered June, July, and August; and fall contained measurements from September, October, and November.

Projections

All data locations (latitude and longitude) were projected into the universal Transverse Mercator (UTM) coordinate system zone 18N (EPSG 26918), which includes the NE US. Coordinate X values increase for monitors further east while coordinate Y values increase for monitors further north. The default coordinate values of UTM are meters; we converted the coordinates into kilometers for our analysis.

Building the multilevel model

After adding the environmental determinants listed above into a single data frame using the generic "merge" function in the R statistical software, we began building our model. We started by running bivariate models for our outcome by each covariate under consideration. We then built our multilevel model in a stepwise fashion, adding covariates from the smallest (point) to the largest (state) spatial aggregate level. At each progressive model building step, the log likelihood, log likelihood ratio, and AIC values were compared.

To begin our step-wise model building, we first considered the null model:

$$PM_{ijk} = \beta_0 + e_i \quad (4)$$

We then considered the null multilevel model, with county nested within state:

$$PM_{ijk} = \beta_0 + u_{0jk} + e_{ijk} \quad (5)$$

Next, we added the point-level covariates of monitor X and Y coordinates and the power plant buffer to create a multilevel point-covariate model:

$$PM_{ijk} = \beta_0 + \beta_1 Xcoord_{ijk} + \beta_2 Ycoord_{ijk} + \beta_3 PowerPlant_{ijk} + u_{0jk} + e_{ijk} \quad (6)$$

Next, we added the county-level covariates to the multilevel model:

$$PM_{ijk} = \beta_0 + \beta_1 Xcoord_{ijk} + \beta_2 Ycoord_{ijk} + \beta_3 PowerPlant_{ijk} + \beta_4 Population_{jk} + \beta_5 TRI_{jk} + \beta_6 Elevation_{jk} + u_{0jk} + e_{ijk} \quad (7)$$

We then added the state-level covariates to the multilevel model:

$$PM_{ijk} = \beta_0 + \beta_1 Xcoord_{ijk} + \beta_2 Ycoord_{ijk} + \beta_3 PowerPlant_{ijk} + \beta_4 Population_{jk} + \beta_5 TRI_{jk} + \beta_6 Elevation_{jk} + \beta_7 Energy_k + \beta_8 VMT_k + u_{0jk} + e_{ijk} \quad (8)$$

Finally, we added the temporal covariates, creating a multilevel ensemble model:

$$\begin{aligned} PM_{ijk} = \quad & \beta_0 + \beta_1 Xcoord_{ijk} + \beta_2 Ycoord_{ijk} + \beta_3 PowerPlant_{ijk} \\ & + \beta_4 Population_{jk} + \beta_5 TRI_{jk} + \beta_6 Elevation_{jk} \\ & + \beta_7 Energy_k + \beta_8 VMT_k + \beta_9 Season + \beta_{10} Year \\ & + u_{0jk} + e_{ijk} \end{aligned} \quad (9)$$

The multilevel ensemble model was compared to a single level ensemble model to investigate whether the final model should include the random effect of county nested within state, or whether the addition of the county and state specific covariates was sufficient without the random effect:

$$\begin{aligned} PM_{ijk} = \quad & \beta_0 + \beta_1 Xcoord_{ijk} + \beta_2 Ycoord_{ijk} + \beta_3 PowerPlant_{ijk} \\ & + \beta_4 Population_{jk} + \beta_5 TRI_{jk} + \beta_6 Elevation_{jk} \\ & + \beta_7 Energy_k + \beta_8 VMT_k + \beta_9 Season + \beta_{10} Year \\ & + e_{ijk} \end{aligned} \quad (10)$$

In all of the preceding models, the subscripts *i*, *j*, and *k* indicate spatial levels, with point level (level 1) denoted with *i*, county level (level 2) denoted with *j*, and state level (level 3) denoted with *k*. PM_{ijk} is the monthly average $PM_{2.5}$ for monitor *i* in county *j* in state *k*. The random effect u_{0jk} is the effect of county *j* in state *k* on average $PM_{2.5}$. The residual error term, e, reflects the levels of the covariates in each

model, up to all three levels (e_{ijk}).

The final model was determined from this step-wise model building process as the model with the lowest AIC value and with significant log likelihood ratio test statistic compared to the previous step-wise model. The final model contained covariates founds to be significant ($p < 0.05$).

Spatial analysis of multilevel model

To investigate how the addition of covariates impacts the residual spatial variation in our final model, we constructed semivariograms of the residuals from the spatially stepped models, from the null model though the final model. We began the spatial investigation of the final model by constructing a semivariogram of the null model residuals, followed by a residual semivariogram of a model that includes the point-level covariates from the final model, to a model including county-level covariates from the final model, to a model that included state-level covariates from the final model, and finally to the final model which included the temporal variables. The spatial analysis considered models with and without county and state random effects.

By comparing the semivariograms of the residuals at each of these steps, we investigate how the addition of covariates at these spatial levels account for the residual spatial variation compared to the null model. We also fit the final model for each year and for each season separately and constructed semivariograms of the residuals from these models to investigate whether the spatial structure of the residuals is similar across time.

We investigate residual spatial variation through the shape of the residual semivariogram, with a horizontal semivariogram indicating spatial independence (no residual spatial variation) (Cressie 1993, Verdú and García‐Fayos 1998, Mannshardt-Shamseldin, Smith et al. 2010). If a semivariogram of the model residuals is flat, then the error term of the model displays constant variance:

$$e \sim N(0, \sigma^2) \tag{11}$$

Otherwise, the error term of the model varies spatially:

$$e \sim N(0, \sigma^2 \mathbf{R}) \tag{12}$$

where the variance σ^2 is multiplied by a distance matrix **R** to account for the spatial structure of the residuals.

2.3 Results

$PM_{2.5}$

From 2000 – 2014, 32,440 monitors reported average $PM_{2.5}$ to the AQS within our study area. On average, each year in our study included data from 2,162.67 monitors, with the fewest monitors (2,021) reporting in 2006 and the most monitors (2,385) reporting in 2002 (Table 1). We computed a monthly average $PM_{2.5}$ from the reported daily arithmetic means of each monitor in the study area for

180 months (2000 – 2014), totaling 32,440 monitor-months in our study. This total included data from 265 monitor sites in 134 counties in 14 states. The average number of daily observations that generated the monthly average $PM_{2.5}$ was 20.56 (range 1 – 692; median 10). On average, PM monitors in our study were located 399.60 km apart (range 0 – 1,652.42 km).

The mean monthly average $PM_{2.5}$ in the NE US from 2000 – 2014 was 11.22 $\mu g/m^3$ (Median = 10.59, Range 1.17 – 47.52 $\mu g/m^3$) (Table 1). A site called "Edgewood" in Harford County, Maryland reported the highest monthly average $PM_{2.5}$ value in our dataset, for July 2002, with an average $PM_{2.5}$ = 47.52 $\mu g/m^3$. This monthly average was based on five measurements taken that month. An unnamed site in New Castle, Delaware, reported the second highest average $PM_{2.5}$ value in our dataset, with an average $PM_{2.5}$ = 43.17 $\mu g/m^3$ in July 2002, based on ten measurements taken that month. Notably, measurements taken in July of 2002 accounted for the top eight average $PM_{2.5}$ concentrations in our study, with averages ranging from 37.61 $\mu g/m^3$ in Northampton County, Pennsylvania to the aforementioned maximum of 47.52 $\mu g/m^3$ in Harford County, Maryland. This period coincides with heavy wildfire activity in Quebec, Canada, and the long range transport of fire particulates may have contributed to these peak $PM_{2.5}$ concentrations (DeBell, Talbot et al. 2004, Le, Breysse et al. 2014).

We calculated an average $PM_{2.5}$ of -1.33 $\mu g/m^3$ for site US-EPA Laboratory in Washington County, Rhode Island, in May 2014. This negative monthly average $PM_{2.5}$ pulled from 62 daily arithmetic mean values, with a standard deviation of 2.04 $\mu g/m^3$. Based on the definition of the daily arithmetic mean by EPA noted previously,

a negative monthly average is nonsensical. Investigating the original file from the AQS database, we found that 48 of the 62 observations for that site and month listed negative arithmetic mean daily values. These data points were removed from the dataset, so that the monthly average $PM_{2.5}$ for site US-EPA Laboratory in Washington County, Rhode Island, was 1.71 μg/m^3 for May 2014, based on the 14 non-negative daily arithmetic mean values, with a standard deviation of 1.10 μg/m^3.

Environmental determinants

Power plants

The AIC values and ANOVA p-values for the comparison of different buffer values are reported in Table 2. ANOVA p-values indicate that models differed significantly as larger buffer sizes were considered from 0 km (no buffer) to 50 km around each power plant location, but no significant difference was detected comparing models with 50 km to models with 100 km buffers (p = 1.000, Table 2). The 10 km power plant buffer was selected for inclusion in further modeling despite having a slightly higher AIC value than the 5 km buffer in our comparative buffer analysis (753.11 vs. 747.75, Table 2) to account for the potentially larger geographic reach of precursor pollutants oxidizing into secondary $PM_{2.5}$ (Paciorek, Yanosky et al. 2009, Hand, Schichtel et al. 2012, De Gouw, Parrish et al. 2014).

Of the 32,440 $PM_{2.5}$ monitors included in our study, 39.96% (12,948) lay within 10 km of a power plant (Table 1).

Toxics release inventory

The mean annual total industrial releases for our study area from 2000 – 2014 was 1,109,653.46 pounds (range 0 – 41,873,902.63 pounds) (Table 1). The county TRI quintiles were defined by total releases in pounds as follows:

- Quintile 1: Percentile range (0, 20.00); Releases range (0 – 12,039.00 pounds)
- Quintile 2: Percentile range (20.10, 40.00); Releases range (12,045.12 – 76,666.90 pounds)
- Quintile 3: Percentile range (40.10, 60.00); Releases range (76,736.51 – 320,203.52 pounds)
- Quintile 4: Percentile range (60.10, 80.00); Releases range (321,155.00 – 1,347,624.00 pounds)
- Quintile 5: Percentile range (80.10, 100); Releases range (1,348,016.48 – 41,873,902.63 pounds)

Season

From 2000 – 2014, $PM_{2.5}$ measurements were fairly evenly distributed by season, with 8,092 (24.98%) in winter, 8,082 (24.94%) in spring, 8,117 (25.05%) in summer, and 8,149 (25.15%) in fall (Table 1).

Other environmental determinants

Descriptive statistics for population, elevation, energy, and traffic are reported in Table 1. Further exploratory data analyses can be found in Appendix A of this dissertation document.

Multilevel Model

Table 3 summarizes the bivariate model results and the ensemble model results (a multilevel model utilizing all of the environmental determinants, formula 10).

The step-wise model building results using the ensemble model environmental determinants are summarized in Table 4. Moving from the null single-level (formula 4) to the null multilevel model (formula 5), the log likelihood ratio test statistic (D) is much greater than the critical value of 5.991 from the chi^2 distribution table for alpha = 0.05 and 2 degrees of freedom (Appendix A), suggesting there is strong evidence for county nested within state effects on $PM_{2.5}$ and supporting the use of a multilevel model. The AIC value is also lower in the multilevel model, further supporting this choice over the single level model.

Comparing the point-level model (formula 6) to the multilevel null model (formula 5), the test statistic is again greater than the critical value ($D > chi^2_{0.05}$, df = 3) and the AIC value is lower for the multilevel point model compared to the multilevel null model (Table 4). When we add the county-level covariates, the test statistic D of the multilevel county-covariate model (formula 7) compared to the point-covariate model is again greater than the critical value ($D > chi^2_{0.05}$, df = 6) and the AIC value is lower for the model that includes county level covariates. Adding in the state-level covariates (formula 8) further reduces the AIC value and also returns a significant test statistic ($D > chi^2_{0.05}$, df = 2), as does adding the temporal covariates to create the ensemble model (formula 9). Finally, a comparison of the ensemble model with (formula 9) and without (formula 10) the random effects of county

nested within state shows that the multilevel ensemble model was significantly improved over the single level model ($D > chi^2_{0.05}$, df = 2) and the AIC value was lower for the multilevel model. Therefore, our final model is a multilevel model with counties nested within states and included environmental covariates at the point, county, and state levels.

We ran the ensemble multilevel model (formula 9) to determine covariate significance ($p < 0.05$), which resulted in the removal of the following covariates from the final model: the point-level X coordinate of the monitor, the county-level TRI quintile, and the county-level population variable (Table 3). The significant environmental determinants from the ensemble model (monitor Y coordinate, power plant buffer, county-level elevation, state-level energy generation, state-level traffic, season, and year) are included in the final model (Table 5). Thus, the final multilevel model is:

$$\begin{aligned} PM_{ijk} = \quad & \beta_0 + \beta_2 Ycoord_{ijk} + \beta_3 PowerPlant_{ijk} + \beta_6 Elevation_{jk} \\ & + \beta_7 Energy_k + \beta_8 VMT_k + \beta_9 Season + \beta_{10} Year \\ & + u_{0jk} + e_{ijk} \end{aligned} \quad (13)$$

Spatial analysis of multilevel model

Figure 2 shows the weighted least squares fit line of the residual semivariograms of the step-wise spatial investigation of the final model, from the

null model (formula 4) to a model including point-level covariates of the final model with no random effect (a single level model, SLM):

$$PM_{ijk} = \beta_0 + \beta_2 Ycoord_i + \beta_3 PowerPlant_i + e_i \quad (14)$$

to a model including the county-level covariate without a county-level random effect (county SLM):

$$PM_{ijk} = \beta_0 + \beta_2 Ycoord_i + \beta_3 PowerPlant_i + \beta_6 Elevation_j + e_i \quad (15)$$

to a model that adds the county-level random effect (county multilevel model, MLM):

$$PM_{ijk} = \beta_0 + \beta_2 Ycoord_{ij} + \beta_3 PowerPlant_{ij} + \beta_6 Elevation_j + u_{0j} + e_{ij} \quad (16)$$

and finally to a model that included state-level covariates, first without a state level random effect:

$$PM_{ijk} = \beta_0 + \beta_2 Ycoord_{ij} + \beta_3 PowerPlant_{ij} + \beta_6 Elevation_j + \beta_7 Energy_k + \beta_8 VMT_k + u_{0j} + e_{ij} \quad (17)$$

Then adding nested random effect, with county nested within state:

$$PM_{ijk} = \beta_0 + \beta_2 Ycoord_{ijk} + \beta_3 PowerPlant_{ijk} + \beta_6 Elevation_{jk} + \beta_7 Energy_k + \beta_8 VMT_k + u_{0jk} + e_{ijk} \quad (18)$$

The semivariograms approach a straight line as we add our covariates and random effects, indicating that the covariates in the final model account for the residual spatial variation of the data (Figure 2). The semivariance differs by model as evidenced by the Y-axis intercepts of Figure 2. This indicates differing variability in the residuals for the models. The multilevel model output indicates that the between-state variance is higher than the between-monitor variance and the between-county variance, and the higher Y intercepts of the state covariate and state random intercept models supports this finding (lower right quadrant, Figure 2). While the residuals show increased variance in the state-level models, the addition of the state-level covariates and random intercept did account for the residual spatial variation seen in earlier models (other quadrants, Figure 2), and thus will be retained in the final model.

Figure 3 displays the weighted least squares line fit of the residual semivariogram of the final model (formula 13). Again, we see a change in the Y intercept of the semivariogram compared to previous models (Figure 2); however, the line is flat, indicating that the inclusion of the covariates and the multilevel structure of our final model accounts for the spatial autocorrelation of the $PM_{2.5}$ data.

Therefore, we retain the covariates and the random effects, and the error term of our final model displays constant variance:

$$\begin{aligned} PM_{ijk} = \quad & \beta_0 + \beta_2 Ycoord_{ijk} + \beta_3 PowerPlant_{ijk} + \beta_6 Elevation_{jk} \\ & + \beta_7 Energy_k + \beta_8 VMT_k + \beta_9 Season + \beta_{10} Year \\ & + u_{0jk} + e_{ijk} \end{aligned} \tag{13}$$

where

$$e_{ijk} \sim N(0, \sigma^2) \tag{19}$$

2.4 Discussion

Our analysis identified significant environmental determinants of $PM_{2.5}$ pollution and accounted for the spatial and temporal influences in building our final model. We considered environmental covariates at different spatial aggregate levels, from point through state level. The inclusion of small scale (point and county level) covariates acknowledges the influence of geographically proximate sources on pollution levels registered by a PM monitor, while the inclusion of the large scale (state level) covariates recognizes the geographic reach of $PM_{2.5}$ and its constituents. The significance of the state level covariates is consistent with the literature; Saunders and Waugh (2015) noted that variability in average daily mean $PM_{2.5}$ was similar across states in the NE US and concluded that the majority of high $PM_{2.5}$ events have an influence over a large spatial area.

Power plant emissions have been shown to be an important consideration in modeling PM trends in the United States (Hand, Schichtel et al. 2012). We use the point-level power plant locations and the state-level net energy production variables to account for the impact of the energy sector on $PM_{2.5}$ concentrations at various scales. Since it is well established that the electricity generation sector is a major contributor to $PM_{2.5}$ pollution in the NE US, it is not surprising that both of these covariates were significant in our final model (Marufu, Taubman et al. 2004, Paciorek, Yanosky et al. 2009, Hand, Schichtel et al. 2012). Mobile sources, and particularly vehicle traffic, is also a well established contributor to $PM_{2.5}$ pollution, and thus the significance of the traffic covariate in the final model was also expected (Paciorek, Yanosky et al. 2009, United States Environmental Protection Agency 2016). The significant negative association between elevation and $PM_{2.5}$ in the final model also follows a previous study (Silcox, Kelly et al. 2012). The retention of the monitor Y coordinate in our final model indicates a significant negative association between monitors in the northern parts of the NE US and the $PM_{2.5}$ data. This agrees with the spatial exploratory data analysis (Appendix A). The final model also retains the temporal variables; the influence of season on $PM_{2.5}$ pollution is well established, as is the overall decreasing trend of $PM_{2.5}$ by year (Malm, Schichtel et al. 2004, Saunders and Waugh 2015, United States Environmental Protection Agency 2016).

Limitations

Our research relies on previously collected data, and our conclusions are limited by the methods and reporting of the primary data collections. Furthermore, data was not available for every year of our study for all environmental

determinants. For the covariates of population, traffic, and energy generation, we utilized existing data to estimate the values for missing years as described above.

We assumed that locations within our study area that do not have a $PM_{2.5}$ monitor in the AQS system (unsampled locations) are "missing at random", with no pattern to the missing pollution data. Visual inspections of the data on the map using GIS show that there may be trends to the monitor placement that could impact our outcomes. For example, if monitors are routinely placed away from $PM_{2.5}$ sources such as power plants or high trafficked roads, our estimates could be biased. This potential bias warrants further investigation. One example of monitor location trends is the relative abundance of $PM_{2.5}$ monitors in urban versus rural locations (Appendix A). However, Hand, Schichtel et al. (2012) noted that urban and rural sites showed similar trends in $PM_{2.5}$ sulfate particles across the US in 2000 – 2010 despite the trend of heavier monitor placement in urban sites.

For computational reasons, we did not include meteorological covariates such as relative humidity and temperature in our study. However, general meteorological trends are represented in our covariates of season, elevation, and latitude, and annual meteorological trends may be represented by our inclusion of year in our study.

Strengths

A unique aspect of our research is our reliance on publicly available data from multiple US federal agencies to arrive at a best-fitting model for $PM_{2.5}$ analysis. We arrived at a final model that is computationally efficient, and we utilized free, open source R statistical software and QGIS geographic information system to

complete our analyses. The accessibility of this research may lend its application to resource-limited agencies and researchers who may use similar techniques to investigate air pollution exposure in epidemiology studies (Kloog, Chudnovsky et al. 2014).

Our recognition of the spatial nature of $PM_{2.5}$ strengthens our final model. By testing for residual spatial variation, we assure that our model acknowledges the spatial autocorrelation of our outcome data. The inferences based on our final model thus avoids the potential for spurious effects as seen in models that fail to account for residual spatial variation (Cressie 1993, Schabenberger and Gotway 2005, Bivand, Pebesma et al. 2008, Berman, Breysse et al. 2015).

2.5 Conclusion

We undertook a deliberate, step-wise approach to build our final model, comparing model performance and significance of covariates as well as investigating residual spatial dependence, to arrive at a statistically sound working model to describe $PM_{2.5}$ in the NE US from 2000 - 2014. In utilizing publicly available data and testing for the effectiveness of every addition to our model, we maximized the feasibility of the model, ensuring a parsimonious model with the fewest parameters that adequately explain the outcome variable. Our analysis concluded that the environmental determinants of monitor Y coordinate, power plant location, elevation, energy generation, traffic, season, and year adequately explained the spatial variation in $PM_{2.5}$ in the NE US from 2000 – 2014. Thus, it is appropriate to use a multilevel model with a constant variance (error term) rather than a kriged

model with a non-constant variance to further explore the data. Our final multilevel model will be utilized in further explorations of this dataset in subsequent chapters of this dissertation document.

2.6 Tables and Figures

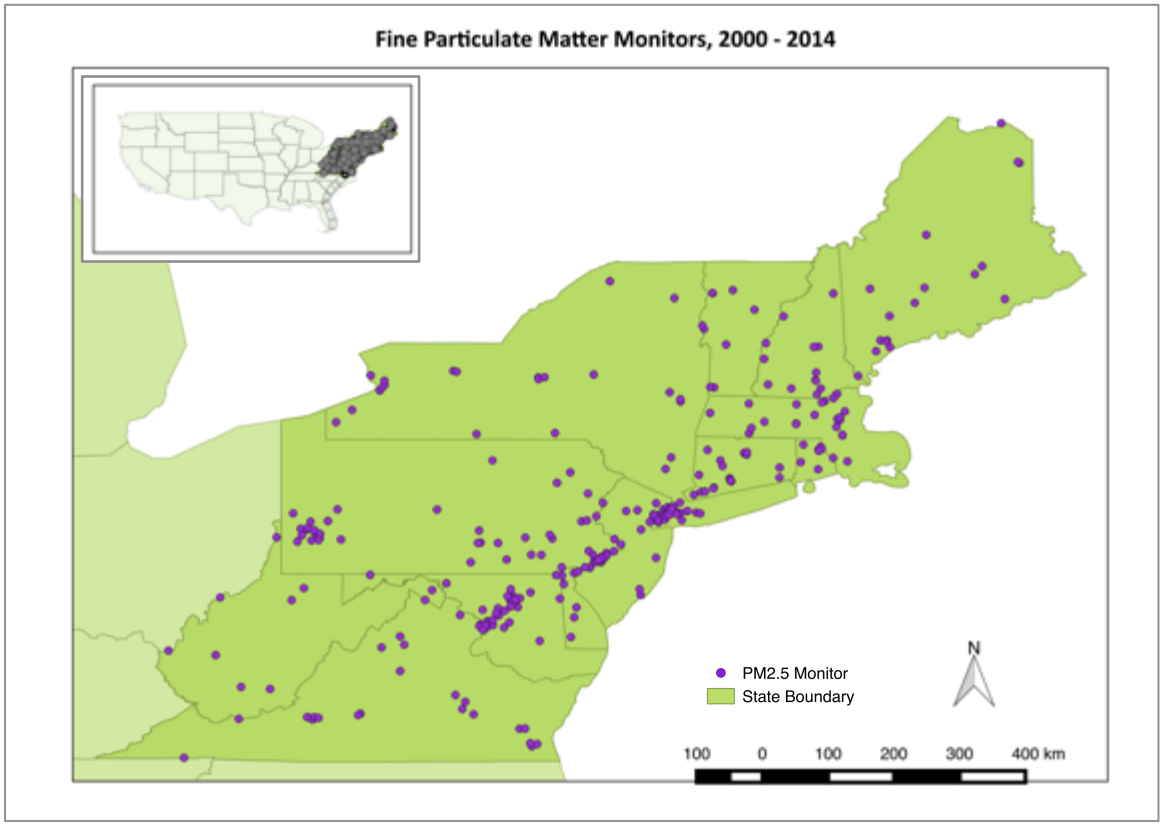

Figure 1. Map of Northeastern United States with EPA AQS monitor locations, 2000 – 2014.

	Variable	Description	Mean	Median	Range
Point level	PM2.5	Outcome variable	11.330	10.680	(1.171, 47.502)
	X km	Monitor X coordinate (UTM zone 18N)	506.100	527.300	(-148.600, 1037.000)
	Y km	Monitor Y coordinate (UTM zone 18N)	4542.000	4509.000	(4138.000, 5266.000)
County level	TRI total releases	Total releases of all chemicals (pounds)	1109653.456	154559.470	(0, 41873902.630)
	Population density	1000 people per county square mile	5.925	2.151	(0.000, 74.000)
	Elevation	County mean in meters	121.600	69.220	(5.917, 740.600)
State level	Traffic density	Billions of miles traveled per state square mile	11.020	12.400	(0.487, 27.730)
	Energy generation	Hundred thousand MWH net energy generation per state square m	11.050	10.920	(0.124, 28.040)
Temporal	Year	Number of monitors per year in study	2163	2145	(2021, 2385)
Temporal	Season	Number of monitors per season in study	Fall	8149	
			Spring	8082	
			Summer	8117	
			Winter	8092	
Point level	Power plant buffer (10 km)	Number monitors inside buffer | percent	12948	39.963	
		Number monitors outside buffer | percent	19492	60.160	

Table 1. Summary of outcome variable and covariates considered in analysis. The monitor coordinate values (X km and Y km) correspond to the universal Transverse Mercator map projection (UTM zone 18N). Coordinate X values increase for monitors further east while coordinate Y values increase for monitors further north.

Buffer size	AIC	ANOVA p-value
No buffer	762.040	
5 km	747.750	0.000
10 km	753.110	0.000
50 km	764.030	0.000
100 km	763.890	1.000*

* Not significant

Table 2. Final model goodness of fit analyses using various buffer sizes around power plant locations. For each buffer size, analysis of variance (ANOVA) p-value compares the model using that buffer size to the model using the buffer size listed directly above it.

	Covariate	Bivariate Models		Ensemble Model		
		Unadjusted Beta hat	SE	Fixed effects Beta hat	SE	p-value
Point level	Monitor X coordinate (X km)	-4.80E-06	0.0001	-0.0008	0.0011	0.4504*
	Monitor Y coordinate (Y km)	-0.0059	0.0001	-0.0030	0.0012	0.0123
	Power Plant 10km buffer	1.2950	0.0475	0.7588	0.0640	0.0000
County level	Population (1000 per year / mi2)	0.0155	0.0025	0.0174	0.0109	0.1115*
	TRI level 1 (reference)					
	TRI level 2	0.5008	0.1167	0.1428	0.1392	0.3051*
	TRI level 3	-0.1720	0.1093*	0.0168	0.1611	0.9169*
	TRI level 4	0.6325	0.1043	-0.1945	0.1699	0.2524*
	TRI level 5	2.4702	0.1030	0.0070	0.1861	0.9699*
	Elevation	-0.0010	0.0002	-0.0035	0.0010	0.0009
State level	Energy (100K MWH per mo / mi2)	0.1726	0.0032	0.2561	0.0111	0.0000
	Traffic (Billions miles per year / mi2)	0.0404	0.0037	0.3708	0.0471	0.0000
Time	Season: Fall (reference)					
	Season: Spring	-0.1875	0.0608	0.0316	0.0484	0.5132*
	Season: Summer	4.1865	0.0606	3.6785	0.0530	0.0000
	Season: Winter	1.6944	0.0608	1.7466	0.0490	0.0000
	Year	-0.4196	0.0049	-0.4190	0.0046	0.0000

* Not significant

Table 3. Summary of bivariate models and the ensemble model results. The ensemble model is a multilevel model utilizing all of the environmental determinants. Significance is determined at $\alpha = 0.05$.

Model	AIC	Log liklihood (LL)	LL ratio test statistic (D)	D df	$chi^2_{0.05}$
Null SLM	162403.500	-98475.230			
Null MLM	156537.100	-78264.540	40421.380	2	5.991
Point MLM	156496.600	-78241.280	46.520	3	7.815
County MLM	155187.600	-77580.820	1320.920	6	12.592
State MLM	151891.700	-75930.850	3299.940	2	5.991
Ensemble MLM	140593.400	-70277.710	11306.280	4	9.488
Ensemble SLM	142992.200	-71479.110	2402.800	2	5.991

Table 4. Summary of the step-wise model building results, moving from a null single level model (SLM) through the ensemble multilevel model (MLM). For each model, the log likelihood ratio test statistic (D) compares the model to the model listed directly above it. Significance is determined by comparing D to the chi^2 critical value ($chi^2_{0.05}$, degrees of freedom = df); all D are significant ($D > chi^2_{0.05}$, df).

	Covariate	Final Multilevel Model		
		Fixed effects Beta hat	SE	p-value
Point level	Monitor Y coordinate (Y km)	-0.0034	0.0012	0.0051
	Power Plant 10km buffer	0.7453	0.0637	0.0000
County level	Elevation	-0.0031	0.0009	0.0011
State level	Energy (100K MWH per mo / mi2)	0.2554	0.0111	0.0000
	Traffic (Billions miles per year / mi2)	0.3587	0.0467	0.0000
Time	Season: Fall (reference)			
	Season: Spring	0.0315	0.0484	0.5154*
	Season: Summer	3.6799	0.0530	0.0000
	Season: Winter	1.7472	0.0490	0.0000
	Year	-0.4167	0.0041	0.0000

* Not significant

Table 5. Summary of final model results. The multilevel final model retains the significant covariates from the ensemble model. Significance is determined at $\alpha = 0.05$.

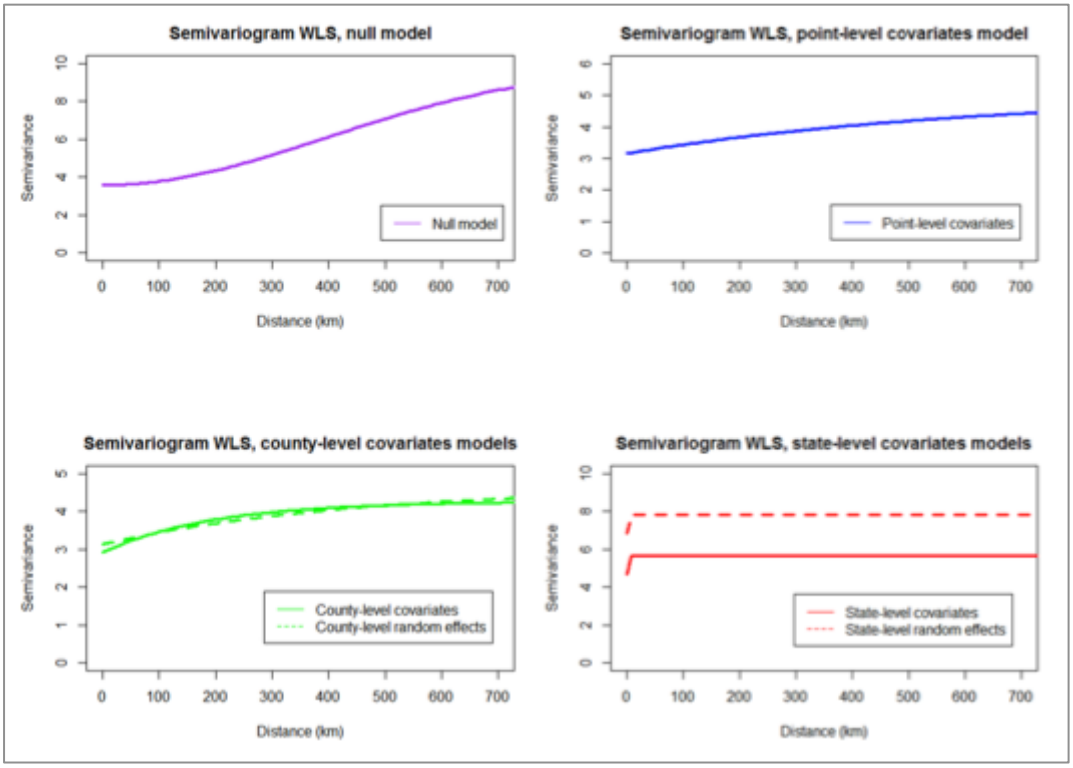

Figure 2. Weighted least squares (WLS) lines of residual semivariograms for step-wise models, from the null model through a model that includes the point-, county-, and state-level covariates. Semivariograms approach a flat line as residuals display spatial independence. The models using all of the covariates (point, county, and state), displayed in the lower right quadrant ("Semivariogram WLS, state-level covariates models"), indicates that the inclusion of the covariates accounts for the residual spatial dependence displayed in earlier models (the null, point-level, and county-level covariates models).

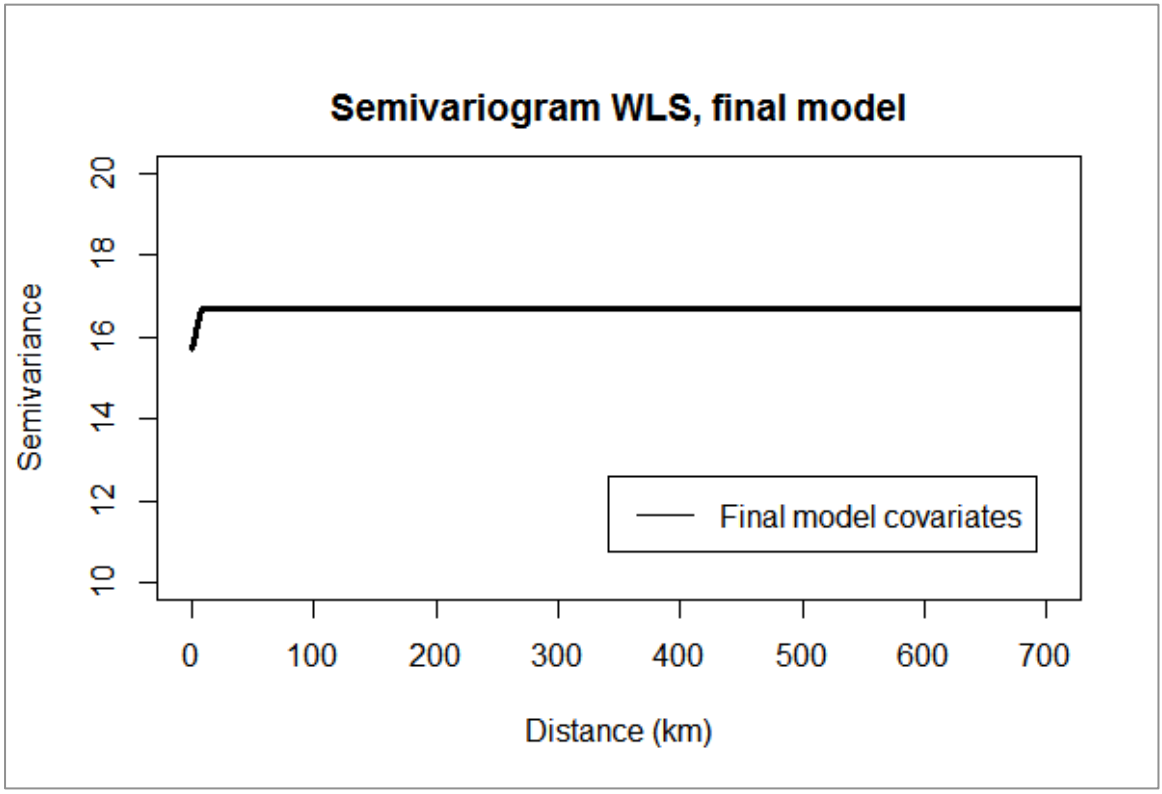

Figure 3. Weighted least squares (WLS) line of the final model. The flat line indicates that the inclusion of the final model covariates accounts for the spatial autocorrelation of $PM_{2.5}$ in the NE US, 2000 - 2014.

2.7 References

Berman, J. D., et al. (2015). "Evaluating methods for spatial mapping: Applications for estimating ozone concentrations across the contiguous United States." Environmental Technology & Innovation **3**: 1-10.

Bivand, R. S., et al. (2008). "Applied spatial data analysis with R. Springer."

Burnham, K. P. and D. R. Anderson (2004). "Multimodel inference understanding AIC and BIC in model selection." Sociological methods & research **33**(2): 261-304.

Cressie, N. (1993). "Statistics for spatial data: Wiley series in probability and statistics." Wiley-Interscience New York **15**: 16.

De Gouw, J., et al. (2014). "Reduced emissions of CO2, NOx, and SO2 from US power plants owing to switch from coal to natural gas with combined cycle technology." Earth's Future **2**(2): 75-82.

DeBell, L. J., et al. (2004). "A major regional air pollution event in the northeastern United States caused by extensive forest fires in Quebec, Canada." Journal of Geophysical Research: Atmospheres **109**(D19).

Federal Highway Administration (2015). Annual Highway Statistics: Roadway Extent, Characterisitcs and Performance. Office of Highway Policy Information.

Gillies, J. A. and A. W. Gertler (2000). "Comparison and evaluation of chemically speciated mobile source PM2. 5 particulate matter profiles." Journal of the Air & Waste Management Association **50**(8): 1459-1480.

Greco, S. L., et al. (2007). "Spatial patterns of mobile source particulate matter emissions-to-exposure relationships across the United States." Atmospheric Environment **41**(5): 1011-1025.

Hand, J., et al. (2012). "Particulate sulfate ion concentration and SO 2 emission trends in the United States from the early 1990s through 2010." Atmospheric Chemistry and Physics **12**(21): 10353-10365.

Hart, J. E., et al. (2009). "Spatial Modeling of PM(10) and NO(2) in the Continental United States, 1985–2000." Environmental Health Perspectives **117**(11): 1690-1696.

Hasheminassab, S., et al. (2014). "Long-term source apportionment of ambient fine particulate matter (PM 2.5) in the Los Angeles Basin: A focus on emissions reduction from vehicular sources." Environmental Pollution **193**: 54-64.

Indiana Department of Environmental Management (2015). "CAIR, Cross-State Air Pollution Rule (CSAPR) and Transport Rule Timeline." Retrieved 3/13/2016, from http://www.in.gov/idem/airquality/2557.htm.

Kim, P., et al. (2015). "Sources, seasonality, and trends of southeast US aerosol: an integrated analysis of surface, aircraft, and satellite observations with the GEOS-Chem chemical transport model." Atmospheric Chemistry and Physics **15**(18): 10411-10433.

Kloog, I., et al. (2014). "A new hybrid spatio-temporal model for estimating daily multi-year PM 2.5 concentrations across northeastern USA using high resolution aerosol optical depth data." Atmospheric Environment **95**: 581-590.

Laden, F., et al. (2000). "Association of fine particulate matter from different sources with daily mortality in six US cities." Environmental Health Perspectives **108**(10): 941.

Le, G. E., et al. (2014). "Canadian forest fires and the effects of long-range transboundary air pollution on hospitalizations among the elderly." ISPRS International Journal of Geo-Information **3**(2): 713-731.

Levy, J. I., et al. (2009). "Uncertainty and Variability in Health-Related Damages from Coal-Fired Power Plants in the United States." Risk Analysis: An International Journal **29**(7): 1000-1014.

Malm, W. C., et al. (2004). "Spatial and monthly trends in speciated fine particle concentration in the United States." Journal of Geophysical Research: Atmospheres **109**(D3).

Mannshardt-Shamseldin, E. C., et al. (2010). "Downscaling extremes: A comparison of extreme value distributions in point-source and gridded precipitation data." The Annals of Applied Statistics: 484-502.

Marufu, L. T., et al. (2004). "The 2003 North American electrical blackout: An accidental experiment in atmospheric chemistry." Geophysical Research Letters **31**(13): L13106.

Miller, P. J. and C. Van Atten (2004). North American power plant air emissions, Comission for Environmental Cooperation of North America.

Paciorek, C. J., et al. (2009). "Practical large-scale spatio-temporal modeling of particulate matter concentrations." Annals of Applied Statistics **3**(1): 370 - 397.

Puentes, R. and A. Tomer (2008). The road… less traveled: An analysis of vehicle miles traveled trends in the US. Metropolitan Infrastructure Initiative Series, Metropolitan Policy Program at Brookings.

Saunders, R. O. and D. W. Waugh (2015). "Variability and potential sources of summer PM 2.5 in the Northeastern United States." Atmospheric Environment **117**: 259-270.

Schabenberger, O. and C. A. Gotway (2005). Statistical methods for spatial data analysis, CRC press.

Silcox, G. D., et al. (2012). "Wintertime PM 2.5 concentrations during persistent, multi-day cold-air pools in a mountain valley." Atmospheric Environment **46**: 17-24.

U.S. Energy Information Administration. "Glossary." Retrieved 4/2/2016, from http://www.eia.gov/tools/glossary/index.cfm?id=n.

U.S. Energy Information Administration (2015). December 2015 Monthly Energy Review. Monthly Energy Review. B. T. Fichman. http://www.eia.gov/totalenergy/data/monthly/.

U.S. Geological Survey (2016, February 12). "Geographic Names Information System ". Retrieved 12/9/2015, from http://geonames.usgs.gov/domestic/download_data.htm.

United States Census Bureau (2010). 2010 Census Summary File 1. American FactFinder.

United States Environmental Protection Agency (2011). AQS Data Dictionary. L. Martin. **2.28:** 425.

United States Environmental Protection Agency (2015, October 27). "1990 Clean Air Act Amendment Summary: Title IV." Retrieved 3/13/2016, from https://www.epa.gov/clean-air-act-overview/1990-clean-air-act-amendment-summary-title-iv.

United States Environmental Protection Agency (2015, December 4). "Air Data." Retrieved 12/7/2015, from https://www3.epa.gov/airquality/airdata/.

United States Environmental Protection Agency (2016, February 21). "Clean Air Interstate Rule (CAIR)." Retrieved 3/13/2016, from https://archive.epa.gov/airmarkets/programs/cair/web/html/index.html.

United States Environmental Protection Agency (2016, February 29). "Cross-State Air Pollution Rule (CSAPR)." Retrieved 2/28/2016, from https://www3.epa.gov/crossstaterule/.

United States Environmental Protection Agency (2016, February 23). "National Trends in Particulate Matter Levels." Retrieved 3/18/2016, from https://www3.epa.gov/airtrends/pm.html.

United States Environmental Protection Agency (2016, February 23). "Particulate Matter (PM)." Retrieved 4/2/2016, from https://www3.epa.gov/airquality/particlepollution/.

United States Environmental Protection Agency (2016, March 29). "Toxics Release Inventory (TRI) Program." Retrieved 4/2/16, from https://www.epa.gov/toxics-release-inventory-tri-program.

Verdú, M. and P. García‐Fayos (1998). "Old‐field colonization by Daphne gnidium: seedling distribution and spatial dependence at different scales." Journal of Vegetation Science **9**(5): 713-718.

CHAPTER 3: AIM 2: INVESTIGATING LARGE SCALE TRENDS AND SMALL SCALE SPATIAL VARIATION IN $PM_{2.5}$ POLLUTION AND THE EFFICACY OF FEDERAL EMISSIONS REGULATIONS IN REDUCING $PM_{2.5}$ POLLUTION IN THE NORTHEASTERN UNITED STATES

Abstract

Research aim 2 investigated the large scale trends and small scale spatial variation in the $PM_{2.5}$ airscape in the NE US from 2000 to 2014. This period marked the passage of federal regulations aimed at reducing $PM_{2.5}$ and $PM_{2.5}$ precursor emissions from two critical sources of particulate pollution in the NE US: power plants and mobile sources including cars and trucks. We examined how the relationship of these sources to $PM_{2.5}$ changed from 2000 to 2014. We utilized innovative methods to assess significant small scale changes in $PM_{2.5}$ pollution across the NE US during this period of changing emission regulations. The analysis showed that while the NE US experienced an overall decrease in $PM_{2.5}$ concentrations from 2000 to 2014, smaller regions within the NE US experienced different trends in the $PM_{2.5}$ airscape during this time. Evidence suggests that regulations aimed at power plant emissions significantly decrease $PM_{2.5}$ pollution.

Abbreviations

AQS	Air quality system
CAA	Clean Air Act
EPA	United States Environmental Protection Agency
CAIR	Clean Air Interstate Rule
CSAPR	Cross-State Air Pollution Rule
NAAQS	National Ambient Air Quality Standards
NE US	Northeast United States
NOx	Nitrogen oxides
PM	Particulate matter
$PM_{2.5}$	Fine particulate matter, < 2.5 μm in aerodynamic diameter
SO_2	Sulfur dioxide

3.1 Introduction

The Clean Air Act (CAA) grants the United States Environmental Protection Agency (EPA) the authority to pass regulations to protect the public's health and welfare from fine particulate matter ($PM_{2.5}$) and other air pollution (United States Environmental Protection Agency 2015). To satisfy this edict, EPA sets national ambient air quality standards (NAAQS) and passes regulations to reduce emissions that contribute to $PM_{2.5}$ air pollution (Figure 1).

Power plants are the dominant source of the $PM_{2.5}$ precursor emissions sulfur dioxide (SO_2) and nitrogen oxides (NOx) (Hand, Schichtel et al. 2012, De Gouw, Parrish et al. 2014). On March 10, 2005, EPA passed the Clean Air Interstate Rule (CAIR) which required fossil fuel fired power plants to reduce their SO_2 emissions over two deadlines (Phase I in 2010 and Phase II in 2015), and also to reduce NOx emissions over deadlines in 2009 (Phase I) and 2015 (Phase II) (Indiana Department of Environmental Management 2015, United States Environmental Protection Agency 2016) (Figure 1). Court challenges to CAIR resulted in a new rule, the Cross-State Air Pollution Rule (CSAPR), replacing CAIR prior to the Phase II deadline in 2015 (United States Environmental Protection Agency 2016).

In addition to power plants, mobile sources also emit $PM_{2.5}$ and precursor emissions (Gillies and Gertler 2000, Greco, Wilson et al. 2007, Hasheminassab, Daher et al. 2014). To control $PM_{2.5}$ and other pollutants from mobile sources, EPA regulates both vehicle emissions and fuel quality. The Tier 2 Motor Vehicle Emissions Standards and Gasoline Sulfur Control Requirements (Tier 2 standards), finalized on February 10, 2000, placed limits on passenger vehicle emissions and

gasoline sulfur content to control pollution (Figure 1) (United States Environmental Protection Agency 2000). Under Tier 2 standards, cars, trucks, and SUV's were required to meet the tailpipe emission standards (0.07 g NOx per mile) beginning with model year 2004 vehicles (United States Environmental Protection Agency 1999). The Tier 2 fuel quality standards required gasoline refiners and importers to cap sulfur levels at 30 ppm average sulfur content, with a maximum sulfur content not to exceed 80 ppm (United States Environmental Protection Agency 1999). These fuel quality standards were phased into effect from 2004 – 2007.

The year 2000 also saw the passage of the Heavy-Duty Engine and Vehicle Standards and Highway Diesel Fuel Sulfur Control Requirements (heavy-duty engine / diesel standards), which set emissions limits for heavy-duty (non-passenger) vehicles and set quality requirements for diesel gasoline (Figure 1) (United States Environmental Protection Agency 2016). The heavy-duty engine / diesel standards required trucks to include a diesel particulate filter as well as NOx exhaust-control technology starting with model years 2007 – 2010, while capping diesel fuel sulfur content at 15 ppm (Manufacturers of Emission Controls Association 2016).

This research investigates the influence of these federal rules on the $PM_{2.5}$ pollution in the northeast United States (NE US) from 2000 to 2014. In Chapter 2 of this dissertation document, a viable spatial regression model for $PM_{2.5}$ pollution was established to identify and quantify environmental determinants that explain variation in $PM_{2.5}$ for the NE US. We use that model to investigate the dynamic influence of environmental determinants on $PM_{2.5}$ pollution in the NE US and how

this relationship varies with the introduction of federal regulations over time. We then investigate the small scale changes in $PM_{2.5}$ pollution from 2000 to 2014 by spatially predicting $PM_{2.5}$ concentrations in unsampled locations (areas without a $PM_{2.5}$ monitor in the EPA's Air Quality System (AQS) network) and develop an approach for statistically comparing two "airscape" maps.

EPA describes decreasing $PM_{2.5}$ concentration trends across the US since 2000, and Saunders and Waugh (2015) affirms this trend in the NE US (United States Environmental Protection Agency 2016). We hypothesize that our research will reaffirm this large scale trend; however, we anticipate that our investigation will identify smaller scale variations in the NE US, including regions within the NE US that did not experience a significant decrease in $PM_{2.5}$ concentrations from 2000 to 2014 despite the large scale trend.

3.2 Methods

Study Area

The study area mirrors the study boundaries described in chapter 2 of this dissertation document (Figure 2). We define the NE US as encompassing the following 14 states: Connecticut, Delaware, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, Virginia, Washington D.C., and West Virginia (Saunders and Waugh 2015).

$PM_{2.5}$

We downloaded daily summary $PM_{2.5}$ concentrations (Code 88101) for each year (2000 – 2014) from the EPA AQS website (United States Environmental Protection Agency 2015). The AQS network of air monitors extends across the United States. For our analysis, we considered the daily $PM_{2.5}$ data for all monitors located in our study area. For each monitor in our study, we averaged the daily data into a monthly average $PM_{2.5}$ value for that monitor.

Federal Regulations

We considered EPA regulations that could alter the $PM_{2.5}$ airscape during our study period of 2000 – 2014 (Figure 1). We focused on the three EPA rules implemented during this time that impact $PM_{2.5}$ and precursor emissions from mobile sources and from power plants: Tier 2 standards and heavy-duty engine / diesel standards, which influence emissions and fuel quality for passenger and non-passenger vehicles (collectively referenced as "mobile source standards" in this document), and CAIR, which requires SO_2 and NOx emission reductions from power plants (referred to as "power plant standards" in this document).

Large Scale Trends

The final model established in chapter 2 of this dissertation was used to investigate the dynamic influence of environmental determinants on $PM_{2.5}$ pollution in the NE US:

$$PM_{ijk} = \beta_0 + \beta_1 Ycoord_{ijk} + \beta_2 PowerPlant_{ijk} + \beta_3 Elevation_{jk} + \beta_4 Energy_k + \beta_5 VMT_k + \beta_6 Season + \beta_7 Year + u_{0jk} + e_{ijk} \quad (1)$$

in which the subscripts *i*, *j*, and *k* indicate spatial levels, with point level (level 1) denoted with *i*, county level (level 2) denoted with *j*, and state level (level 3) denoted with *k*. PM_{ijk} is the monthly average $PM_{2.5}$ for monitor *i* in county *j* in state *k*. $Ycoord_{ijk}$ is the Y coordinate value of the PM monitor *i* in county *j* in state *k*, corresponding to the universal Transverse Mercator map projection (UTM zone 18N) and expressed in km. $PowerPlant_{ijk}$ is a binary variable, indicating whether monitor *i* in county *j* in state *k* lies within 10 km of a power plant, and is equal to 1 if monitor $_i$ is located within 10 km of a power plant and equal to 0 otherwise. $Elevation_{jk}$ is the county-average elevation. $Energy_k$ is the state-level monthly net energy generation, expressed as hundred thousand megawatt hours (MWH) per state square mile. VMT_k is the traffic density covariate of annual vehicle miles traveled (VMT) per state, scaled into billion VMT per state square mile. We defined season as follows: winter spanned the months of December, January, and February; spring entailed March, April, and May; summer covered June, July, and August; and fall contained measurements from September, October, and November. The random effect u_{0jk} is the effect of county *j* in state *k* on average $PM_{2.5}$. The residual error term, e_{ijk}, reflects all three levels (point *i*, county *j*, and state *k*), and is assumed to display constant variance following the analysis in chapter 2 of this dissertation document:

$$e_{ijk} \sim N(0, \sigma^2)$$

In investigating the impact of the mobile source and the power plant standards on $PM_{2.5}$ across the study area, we conducted a stratified and a joint analysis. In the stratified analysis, the final model (formula 1) was run on a dataset of $PM_{2.5}$ outcomes and environmental determinants for the first year in the study period (2000). This model represents the relationship between the environmental determinants and the $PM_{2.5}$ outcome prior to the passage of the mobile source standards and the power plant standards. The model was then run on a dataset of $PM_{2.5}$ outcomes and environmental determinants for the last year in the study period (2014), after both the mobile source and the power plants standards were in effect. A comparison of the covariate coefficients allowed us to investigate how the relationship between the covariates (the environmental determinants) and the outcome ($PM_{2.5}$) varied over a time period that included the introduction of these federal regulations. If the covariate coefficients (slopes) are equal but with different regression intercepts, then the regression surfaces for before (2000) and after (2014) the passage of the mobile source and power plant standards are parallel, indicating the relationship between the environmental covariates and the $PM_{2.5}$ outcome did not change (Zimmerman, Liu et al. 1996, Khan 2003). If there is no change in both the covariate coefficients and the regression intercepts, then the regression surfaces for before and after the standards are coincident (Zimmerman, Liu et al. 1996, Weaver and Wuensch 2013).

The joint analysis combined before- and after-standards data and introduced an interaction term into the model. For the joint analysis, $PM_{2.5}$ and environmental determinants data for 2000 and 2014 were combined, and the following model was run:

$$\begin{aligned} PM_{ijk} = \quad & \beta_0 + \beta_1 Ycoord_{ijk} + \beta_2 PowerPlant_{ijk} + \beta_3 Elevation_{jk} \\ & + \beta_4 Energy_k + \beta_5 VMT_k + \beta_6 Season + \beta_7 I_{ijk} \\ & + \beta_8 (I_{ijk} * VMT_k) + \beta_9 (I_{ijk} * PowerPlant_{ijk}) \\ & + \beta_{10} (I_{ijk} * Energy_k) + u_{0jk} + e_{ijk} \end{aligned} \quad (2)$$

where I_{ijk} is an indicator variable for monitor *i* in county *j* in state *k*, equal to 0 for $PM_{2.5}$ measurements taken at that monitor before the introduction of the mobile source and power plant standards (2000) and equal to 1 for measurements taken after the standards came into effect (2014). The interaction term ($I_{ijk} * VMT_k$) allows us to investigate how the relationship between the traffic density covariate VMT_k and the outcome variable PM_{ijk} differs before versus after the implementation of the mobile source standards. The interaction term ($I_{ijk} * PowerPlant_{ijk}$) allows us to investigate how the relationship between monitor proximity to a power plant and the outcome variable PM_{ijk} differs before versus after the implementation of the power plant standards. Similarly, the interaction term ($I_{ijk} * Energy_k$) allows us to investigate how the relationship between the state-level energy model covariate and the outcome PM_{ijk} differs before versus after the implementation of the power plant standards.

The final model (formula 1) fitted to each year (2000 and 2014) was cross validated to assess model performance. Using the sample function of the base package in the R statistical software, a random sample of 10% of the data points was created with shuffling for each year. Formula 1 was used to predict at these locations and the difference between the observed and the predicted values was calculated. The joint analysis model (formula 2) was also cross validated for the joint dataset (2000 and 2014 combined) in the same manner. We report the summary of the differences between the observed and the predicted values for each cross validation to assess model performance.

Small Scale Spatial Variation

To investigate the small scale spatial variation of $PM_{2.5}$ concentrations in the NE US and how these concentrations changed from 2000 to 2014 with the implementation of the mobile source and power plant standards, we developed a geostatistical-based approach for comparing before and after spatial surfaces. This approach extends previously established methods of large scale comparisons to compare spatially paired predicted surfaces in which the spatial pairing is incomplete (Zimmerman, Liu et al. 1996). Application was to the summer months data (June, July, and August) for 2000 (before the passage of the standards) and for 2014 (after the standards) because the summer months showed the greatest influence on the $PM_{2.5}$ concentrations in the NE US among the environmental determinants investigated in chapter 2 of this dissertation document (Chapter 2, Table 5).

The number of AQS monitors and the AQS monitor locations vary from year to year and month to month, yielding incomplete paired spatial designs. To address this incomplete spatial pairing, we first considered the union of monitored locations for a given summer month in 2000 and 2014. This yielded a set of coincident monitored locations between the two years as well as additional locations in each year with missing $PM_{2.5}$ concentration data. We employed ordinary kriging to predict $PM_{2.5}$ concentrations at the set of missing locations for each summer month in 2000 and 2014. Combining these predictions with the monitored data yielded outcomes at coincident monitor locations and thus a complete spatial pairing of summer outcomes in 2000 and 2014. The difference in $PM_{2.5}$ concentrations (2014 concentrations minus 2000 concentrations) was calculated at each monitor location for each summer month, and geostatistical analysis was performed on these differences.

The geostatistical analysis on the $PM_{2.5}$ differences involved descriptive statistics to assess distributional properties and semivariogram estimation to characterize spatial dependence. Parametric semivariogram functional forms were fit to the estimated semivariograms using weighted least squares (Schabenberger and Gotway 2005). Ordinary kriging models were used to generate a spatially predicted surface of the difference in $PM_{2.5}$ with accompanying prediction uncertainties across a grid of the NE US (Figure 2). We used ordinary kriging, a spatial prediction method that is not reliant on covariates, to focus on the small scale variation in $PM_{2.5}$ differences and to complement the separate large scale trend analysis that we performed (Cressie 1990, Cressie 1993). The geographic

information system QGIS (version 2.10.1-Pisa) was used to generate a prediction grid over the study area. The prediction grid contained 3,920 evenly distributed points located 13.194 km apart in 323 counties in all 14 states in the NE US. A map was generated to show the ordinary kriged predicted differences in $PM_{2.5}$ concentrations from 2000 to 2014 for each summer month.

Significance of the spatially predicted surface of the difference in $PM_{2.5}$ was assessed by estimating the probability that each predicted difference was less than zero. A negative (less than zero) difference indicates a decrease in monthly mean $PM_{2.5}$ at a location for a given summer month from 2000 to 2014. We employed conditional simulation to calculate this probability, which simulates data from the kriging predictive distribution (a conditional statistical distribution) while preserving the spatial dependence structure of the measured differences (the semivariogram) (Gotway 1994). Accounting for the spatial dependence structure in the simulations is critical to providing statistically appropriate estimations and interpretations of hypothesis tests for spatially dependent data.

For each summer month, 1,000 conditionally simulated realizations of $PM_{2.5}$ differences were generated at each of the 3,920 prediction grid locations. Each simulated realization is a spatial realization (map) of $PM_{2.5}$ differences. The proportion of these simulated values at each grid location that was negative (less than zero) was calculated, and the p-value defined as 1 minus this proportion. P-values < 0.05 were considered significant. We mapped the p-values to complement the ordinary kriged prediction maps, indicating where $PM_{2.5}$ concentrations differed significantly from 2000 to 2014 for each summer month.

3.3 Results

Large Scale Trends

In 2000, the mean of the monthly average $PM_{2.5}$ in the NE US was 13.42 $\mu g/m^3$ (median = 13.18 $\mu g/m^3$; range 2.10 – 34.31 $\mu g/m^3$). In 2014, the mean was 8.45 $\mu g/m^3$ (median = 8.06 $\mu g/m^3$; range 1.17 – 27.25 $\mu g/m^3$). This difference is both considerable and significant ("Year" covariate $0 < p < 0.00$, Table 1).

The stratified and joint analysis model results are reported in Table 1. The covariates (environmental determinants) most relevant to an exploration of the relative impacts of the mobile sources and power plant standards are the 10 km power plant buffer, the state level traffic density, and the state level net energy generation covariates. From 2000 to 2014, the effect of the proximity of a power plant to a $PM_{2.5}$ monitor decreased slightly ($\beta = 0.90$ in 2000 vs. 0.72 in 2014, Table 1) while the effects of traffic and energy production increased slightly ($\beta = -0.04$ in 2000 vs. -0.09 in 2014 for traffic; $\beta = 0.05$ in 2000 vs. 0.18 in 2014 for energy; Table 1). However, these effects are tempered by the significance of the covariates: state level traffic and energy are not significant predictors of $PM_{2.5}$ concentrations in 2000, but they regain significance in the 2014 stratified analysis ($\alpha = 0.05$).

The joint analysis further explored the power plant buffer, traffic density, and energy generation covariates, and included interaction terms for these covariates with time (Table 1). The interaction between the state level environmental determinants of traffic density and energy generation did not show significant change in the effect of these determinants on $PM_{2.5}$ concentrations;

however, the interaction term for the power plant buffer and year was significant, indicating that the effect of a power plant within 10 km of a $PM_{2.5}$ monitor differs from 2000 to 2014 ($\alpha = 0.05$). The coefficient of the power plant buffer was higher in the joint analysis than in either the 2000 or the 2014 stratified analysis ($\beta = 1.09$ in the joint vs. 0.90 in 2000 and 0.72 in 2014 for energy; Table 1).

The covariate coefficients including the regression intercepts differ between 2000 and 2014, indicating that the models are neither parallel nor coincident. Notably, the spring season coefficient changed effect between the years: in 2000, we estimated a decrease in $PM_{2.5}$ concentrations comparing spring to fall (the reference group) ($\beta = -2.24$, SE = 0.19), while in 2014, spring season showed increased $PM_{2.5}$ concentrations compared to fall ($\beta = 2.29$, SE = 0.13). Furthermore, the covariate coefficient for winter in 2014 is notably greater than the coefficient for summer 2014 ($\beta = 3.08$ in winter vs. 1.16 in summer).

The cross-validation of the final model (formula 1) indicated sufficient agreement between the observed and predicted values for both 2000 and 2014 data. The average difference between the monitor value and the prediction at that monitor was 0.19 $\mu g/m^3$ in 2000 (median = -0.14 $\mu g/m^3$). In 2014, the average difference between the observed and predicted value at each monitor was -0.05 $\mu g/m^3$ (median = -0.14 $\mu g/m^3$). The cross-validation of the joint analysis model (formula 2) also indicated sufficient agreement between observed and predicted values: the average difference was 0.40 $\mu g/m^3$ (median = 0.03 $\mu g/m^3$).

Small Scale Spatial Variation

519 monitors reported $PM_{2.5}$ concentrations in the NE US during the summer of 2000, with 170 reporting in June, 174 in July, and 175 in August. The summer of 2014 entailed 467 $PM_{2.5}$ monitors in the NE US, with 155 reporting in June, 156 in July, and 156 in August. Generating the coincident monitor locations as described previously resulted in 229 coincident monitor locations for June 2000 and 2014, 231 coincident locations for July 2000 and 2014, and 232 coincident locations for August 2000 and 2014.

Figure 3 presents the results of the small scale analysis. The entire NE US experienced a decrease in average monthly $PM_{2.5}$ concentrations comparing June 2000 to June 2014, and this difference was significant in the majority of the study area ($\alpha = 0.05$). The mid-Atlantic region exhibited the greatest decrease in June $PM_{2.5}$ concentrations, specifically the eastern areas of Maryland, Delaware, and New Jersey. The far northeast, including most of Maine as well as northern New Hampshire, Vermont, and northeastern New York, experienced the smallest decrease in June $PM_{2.5}$ concentrations; however, these differences are not significant.

The majority of the NE US also experienced a decrease in July $PM_{2.5}$ concentrations, with the exception of the northeastern section including Maine and northern New Hampshire and Vermont which experienced an increase from 2000 to 2014 (Figure 3). However, the increase in $PM_{2.5}$ concentrations in this northeastern section was not significant. Western Virginia and West Virginia experienced the greatest decrease in July $PM_{2.5}$ concentrations from 2000 to 2014.

Most of the NE US also experienced a decrease in August $PM_{2.5}$ concentrations from 2000 to 2014, with the exception of northern New York and Vermont which showed an insignificant increase in August $PM_{2.5}$ concentrations during this time (Figure 3). Both the eastern coast of the mid-Atlantic states (Maryland, Delaware, and New Jersey) and West Virginia exhibited the greatest decrease in August $PM_{2.5}$ concentrations from 2000 to 2014. Further results from the small scale analyses can be found in Appendix B of this dissertation document.

3.4 Discussion

The decrease in overall mean monthly average $PM_{2.5}$ values from 2000 to 2014 reaffirms the trend of decreasing $PM_{2.5}$ pollution identified across the US by the EPA and in the NE US by Saunders and Waugh (2015) (United States Environmental Protection Agency 2016). The joint analysis results of the large scale trend investigation indicated that the effect of power plant proximity on $PM_{2.5}$ concentrations changed significantly from 2000 to 2014 (Table 1). In 2000, $PM_{2.5}$ monitor proximity to a power plant (the presence of a power plant within 10 km of a monitor) resulted in an average increase of 0.90 $\mu g/m^3$ $PM_{2.5}$ compared to monitors located further than 10 km from power plants and controlling for the monitor Y coordinate, county level elevation, state level traffic density, state level energy generation, and season. In 2014, proximity to a power plant resulted in an average increase of 0.72 $\mu g/m^3$ $PM_{2.5}$, a sizable decrease in effect compared to 2000. In a joint analysis that combined the 2000 and 2014 data, the interaction of year and power plant proximity further supported the conclusion that the effect of power

plant proximity on $PM_{2.5}$ concentrations changed significantly from 2000 to 2014 (Table 1). Thus, the evidence suggests that the passage of the power plant standards between 2000 and 2014 contributed significantly to the trend of decreasing $PM_{2.5}$ pollution in the NE US.

The investigation of the efficacy of the mobile source standards failed to identify a significant change in the effect of traffic density on $PM_{2.5}$ pollution in a joint analysis with an interaction of year and traffic density (Table 1). This may be due to the aggregate level of traffic to the state level in this analysis: a similar lack of significance was observed in the interaction between year and the other state-level environmental determinant, net energy production (Table 1). Further confounding the investigation of the mobile source standards on $PM_{2.5}$ pollution was the apparent negative relationship between traffic density and $PM_{2.5}$ concentrations as identified by the negative slope of the traffic density coefficient in the large scale analyses (Table 1). An inverse relationship between traffic density and $PM_{2.5}$ pollution is inconsistent with the identification of mobile sources as a critical source of $PM_{2.5}$ and precursor emissions (Gillies and Gertler 2000, Greco, Wilson et al. 2007, Hasheminassab, Daher et al. 2014). Further analyses of traffic at smaller levels of spatial aggregation or analyses that include the type of traffic captured by the state level environmental determinant may negate the inverse relationship detected in this analysis. For example, if states with high traffic density also tend to incorporate lower emission vehicles compared to states with low traffic density, then this systematic difference would bias the results of the traffic variable in the model and

impact the conclusions about the efficacy of the mobile source standards in reducing $PM_{2.5}$ pollution.

Motivated by methods established by Gotway (1994) and Zimmerman, Liu et al. (1996), we developed an innovative approach to test for significance that we applied to the small scale spatial variation analysis to determine regions within the NE US that experienced significant change in $PM_{2.5}$ pollution from 2000 to 2014. This analysis demonstrated that different regions within the NE US experienced unique changes the $PM_{2.5}$ airscape from 2000 to 2014, with differences in average summer monthly $PM_{2.5}$ concentrations ranging from -11.24 to 2.92 $\mu g/m^3$ (Figure 3). While the precise areas differ slightly between monthly analyses, the mid-Atlantic and the southwestern states demonstrated the largest decrease in summer $PM_{2.5}$ pollution in the NE US from 2000 to 2014. Notably, the northernmost region of the study area failed to show significant changes in $PM_{2.5}$ pollution from 2000 to 2014. However, this area is sparsely monitored, and the relative lack of data impacted significance testing. States in this sparsely monitored region displayed an increase in July and August $PM_{2.5}$ concentrations, albeit an insignificant one. This increase may be due to high $PM_{2.5}$ events in the area, such as wildfires. In the summer of 2014, Canada experienced its most severe fire season in decades, and the long range transport of particulate pollution from these events may have impacted the northern regions of the NE US (Hand, Schichtel et al. 2012, Veraverbeke 2015).

Limitations

The research presented in this chapter relies on previously collected data and utilizes existing data to estimate values for missing measurements. Thus, the

conclusions are limited both by the methods and reporting of the primary data collections and by the appropriateness of the estimations.

The influence of season identified in this chapter differs from the findings of chapter 2 of this dissertation document, which noted that the summer season displayed the greatest influence on $PM_{2.5}$ concentrations in the NE US when all years are combined in the analysis (Chapter 2, Table 5). In this chapter, we observed that the effect of season changed between 2000 and 2014, and the winter season imparted the greatest impact on $PM_{2.5}$ concentrations in the NE US when these two years are isolated in the analysis. The differences in the season as well as other model covariate coefficients identified in Table 1 indicated that while our final model (formula 1) is appropriate for a large-scale analysis that collapses the $PM_{2.5}$ concentration data temporally for 2000 - 2014, the model may not fit as well to an investigation on individual years. Furthermore, while the model considered the two primary sources of $PM_{2.5}$ in the NE US (power plants and mobile sources), it does not account for high emission events such as wildfires that influence yearly $PM_{2.5}$ concentrations (Saunders and Waugh 2015).

We considered only the EPA federal regulations that introduced stricter power plant and mobile source standards from 2000 to 2014 and assumed that these standards have the greatest impact on the relationships between the associated environmental determinants (power plant proximity, net energy production, traffic density) and $PM_{2.5}$ pollution. However, states and localities may enact stricter rules beyond the limits of the EPA power plant and mobile source standards, which could impact the relationship between these environmental

determinants and $PM_{2.5}$ concentrations in the NE US. EPA can compel a state to enact stricter standards than the federal limits if areas within the state are non-compliant with the NAAQS for $PM_{2.5}$ (Hasheminassab, Daher et al. 2014). Similarly, our analyses did not account for practices of car and truck manufacturers and gasoline refiners beyond the requirements of the EPA standards, which may further influence the relationship between traffic density and $PM_{2.5}$ concentrations in the NE US.

While it is not a limitation of the work described, it is important to note that while this research predicted $PM_{2.5}$ concentrations at unmonitored locations, it did not attempt to estimate human exposure to $PM_{2.5}$ in the study area. Exposure estimation entails a consideration of dose; for example, Gerco et al. (2007) utilized the source-receptor matrix of health risk assessment to investigate the relationship between the mobile sources of $PM_{2.5}$ and $PM_{2.5}$ exposure.

Strengths

This research employed innovative methodologies in both the large and the small scale analyses. The large scale investigation introduced a multilevel regression analysis to assess the relative effects of specific federal policies on $PM_{2.5}$ pollution and concluded that power plant standards focused on reducing $PM_{2.5}$ and precursor emissions have a significant impact on $PM_{2.5}$ pollution. The small scale analysis introduced an innovative approach to test significance that extends the large scale comparison method of Zimmerman, Liu et al. (1996) into a small scale comparison of spatially predicted surfaces and utilized the conditional simulation established by Gotway (1994) to account for and preserve the spatial dependence structure of the

data in the significance test. This allowed us to identify significant small scale changes in $PM_{2.5}$ pollution in a statistically appropriate manner that acknowledges the spatial nature of $PM_{2.5}$. This methodology may be applied in future research to determine whether new regulations attained a target decrease ($\mu g/m^3$) of ambient $PM_{2.5}$ concentrations.

3.5 Conclusion

We implemented analyses of both the large scale and small scale spatial variation of $PM_{2.5}$ pollution in the NE US to test our hypothesis that while the NE US experienced an overall decrease in $PM_{2.5}$ concentrations from 2000 to 2014, different regions within the NE US experienced unique changes in the $PM_{2.5}$ airscape during this time. The small scale analysis supported this hypothesis. The mid-Atlantic and the southwestern regions of the NE US demonstrated the largest decrease in summer $PM_{2.5}$ pollution, while the northernmost region of the NE US failed to show significant changes from 2000 to 2014.

The large scale analysis identified power plant standards as a significant tool in reducing $PM_{2.5}$ pollution across the NE US. This result supports current and future regulations that focus on power plant emission reductions to improve air quality. At the time of this writing, the Cross-State Air Pollution Rule (CSAPR), which replaced the power plant standards investigated in this research (CAIR), has endured legal challenges at both the D.C. Circuit and the U.S. Supreme Court (Indiana Department of Environmental Management 2015, United States Environmental Protection Agency 2016). Future reductions on $PM_{2.5}$ precursor emissions from power plants

are expected from Phase II of CSAPR in 2017 as well as from the Mercury and Air Toxics Standards (MATS), finalized by EPA on December 16, 2011 (United States Environmental Protection Agency 2016). However, MATS has been challenged in court since its inception, and further challenges to CSAPR, MATS, and future power plant regulations are to be expected. Research that investigates the consequences of these regulations can provide critical evidence as EPA defends current rules and considers future regulatory actions. The research presented in this chapter provides scientifically sound evidence that power plant emission controls are effective in reducing $PM_{2.5}$ pollution and supports current rules and future actions that limit power plant emissions to improve air quality.

3.6 Tables and Figures

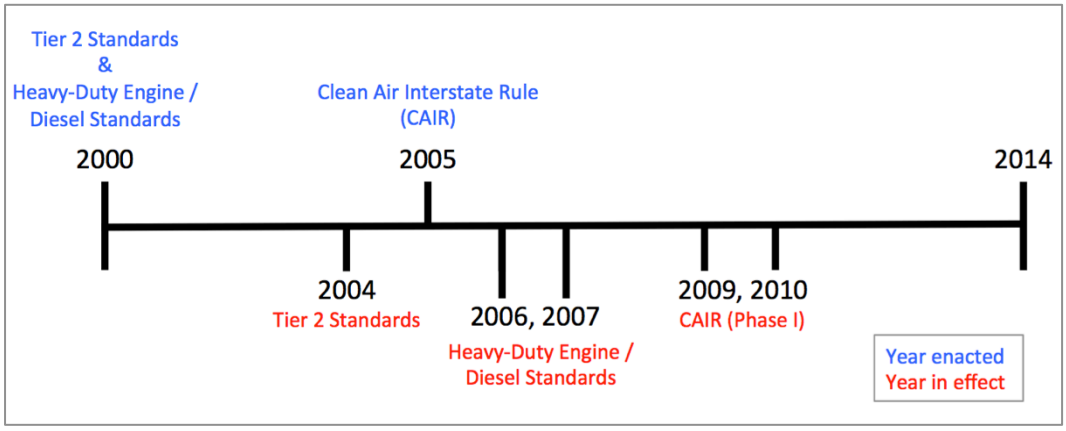

Figure 1. Timeline of EPA regulations aimed at curbing $PM_{2.5}$ and precursor emissions implemented during the study period, 2000 – 2014. The Tier 2 Standards and the Heavy-Duty Engine / Diesel Standards are mobile source standards and the Clean Air Interstate Rule (CAIR) specifies power plant standards.

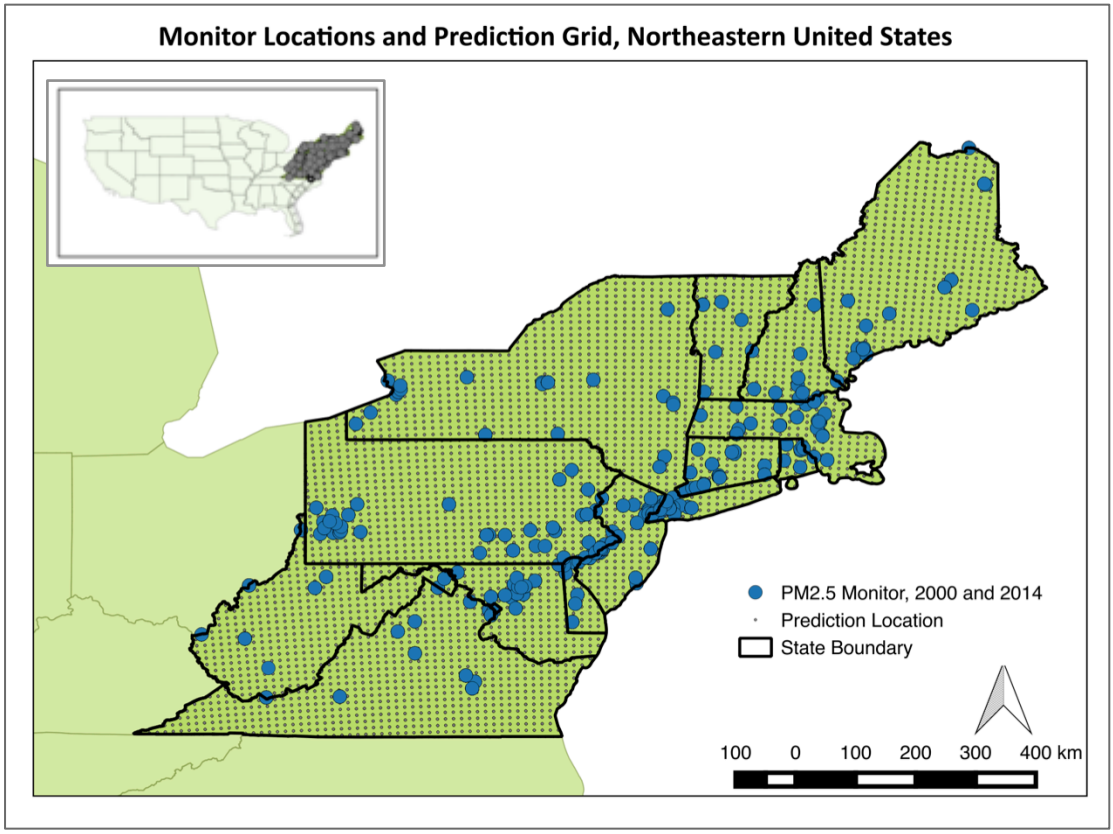

Figure 2. Map of the northeastern United States showing $PM_{2.5}$ monitor locations in 2000 and 2014 and the prediction points.

	Covariate	2000 Stratified Analysis		2014 Stratified Analysis		Joint Analysis, 2000 + 2014		
		Fixed effects Beta hat	SE	Fixed effects Beta hat	SE	Fixed effects Beta hat	SE	p-value
Point level	Power Plant 10km Buffer	0.8978	0.2135	0.7159	0.1601	1.0862	0.1681	0.0000
	Monitor Y Coordinate (Y km)	-0.0085	0.0011	-0.0019	0.0011*	-0.0057	0.0009	0.0000
County level	Elevation	-0.0015	0.0013*	-0.0022	0.0011	-0.0020	0.0010	0.0472
State level	Traffic (Billions miles per year / mi2)	-0.0444	0.0460*	-0.0945	0.0423	-0.0820	0.0381	0.0313
	Energy (100K MWH per mo / mi2)	0.0529	0.0456*	0.1834	0.0366	0.1313	0.0349	0.0002
Time	Season: Fall (reference)							
	Season: Spring	-2.2364	0.1947	0.2875	0.1316	-1.0326	0.1246	0.0000
	Season: Summer	0.6863	0.2111	1.1613	0.1415	0.8782	0.1341	0.0000
	Season: Winter	1.0685	0.1996	3.0817	0.1370	2.0232	0.1281	0.0000
Year, Interactions	Year (2000 vs. 2014)					-4.3436	0.2176	0.0000
	Year * Traffic					-0.0253	0.0186	0.1748*
	Year * Energy					-0.0123	0.0185	0.5080*
	Year * Power Plant Buffer					-0.6121	0.1998	0.0022

* Not significant

Table 1. Summary of stratified and joint analysis model results. Significance is determined at $\alpha = 0.05$ (p-values computed but not shown for stratified analysis).

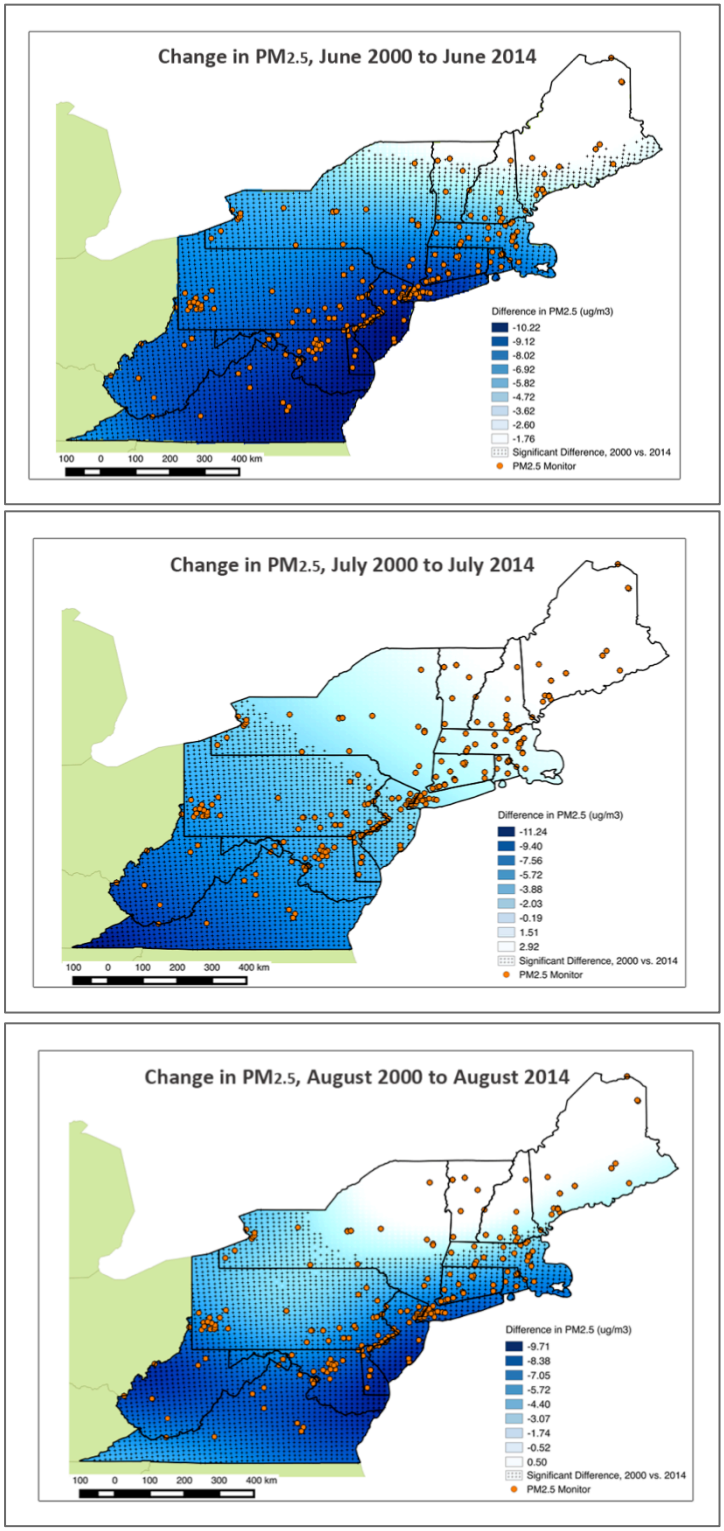

Figure 3. Map of the ordinary kriged predicted differences in $PM_{2.5}$ concentrations from 2000 to 2014 for each summer month. Significance is determined at $\alpha = 0.05$.

3.7 References

Cressie, N. (1990). "The origins of kriging." Mathematical geology **22**(3): 239-252.

Cressie, N. (1993). "Statistics for spatial data: Wiley series in probability and statistics." Wiley-Interscience New York **15**: 16.

De Gouw, J., et al. (2014). "Reduced emissions of CO2, NOx, and SO2 from US power plants owing to switch from coal to natural gas with combined cycle technology." Earth's Future **2**(2): 75-82.

Gillies, J. A. and A. W. Gertler (2000). "Comparison and evaluation of chemically speciated mobile source PM2. 5 particulate matter profiles." Journal of the Air & Waste Management Association **50**(8): 1459-1480.

Gotway, C. A. (1994). "The use of conditional simulation in nuclear-waste-site performance assessment." Technometrics **36**(2): 129-141.

Greco, S. L., et al. (2007). "Spatial patterns of mobile source particulate matter emissions-to-exposure relationships across the United States." Atmospheric Environment **41**(5): 1011-1025.

Hand, J., et al. (2012). "Particulate sulfate ion concentration and SO 2 emission trends in the United States from the early 1990s through 2010." Atmospheric Chemistry and Physics **12**(21): 10353-10365.

Hasheminassab, S., et al. (2014). "Long-term source apportionment of ambient fine particulate matter (PM 2.5) in the Los Angeles Basin: A focus on emissions reduction from vehicular sources." Environmental Pollution **193**: 54-64.

Indiana Department of Environmental Management (2015). "CAIR, Cross-State Air Pollution Rule (CSAPR) and Transport Rule Timeline." Retrieved 3/13/2016, from http://www.in.gov/idem/airquality/2557.htm.

Khan, S. (2003). "Estimation of the parameters of two parallel regression lines under uncertain prior information." Biometrical journal **45**(1): 73-90.

Manufacturers of Emission Controls Association (2016). "U.S. EPA 2007/2010 Heavy-Duty Engine and Vehicle Standards and Highway Diesel Fuel Sulfur Control Requirements." Retrieved 4/12/2016, from http://www.meca.org/regulation/us-epa-20072010-heavyduty-engine-and-vehicle-standards-and-highway-diesel-fuel-sulfur-control-requirements.

Saunders, R. O. and D. W. Waugh (2015). "Variability and potential sources of summer PM 2.5 in the Northeastern United States." Atmospheric Environment **117**: 259-270.

Schabenberger, O. and C. A. Gotway (2005). Statistical methods for spatial data analysis, CRC press.

United States Environmental Protection Agency (1999, December). "Regulatory Announcement: EPA's Program for Cleaner Vehicles and Cleaner Gasoline." Retrieved 4/12/2016, from https://www3.epa.gov/tier2/documents/f99051.pdf.

United States Environmental Protection Agency (2000). Tier 2 Motor Vehicle Emissions Standards and Gasoline Sulfur Control Requirements EPA. Federal Register. **40 CFR Parts 80, 85, and 86**.

United States Environmental Protection Agency (2015, December 4). "Air Data." Retrieved 12/7/2015, from https://www3.epa.gov/airquality/airdata/.

United States Environmental Protection Agency (2015, October 27). "Clean Air Act Overview." Retrieved 3/31/16, from https://www.epa.gov/clean-air-act-overview.

United States Environmental Protection Agency (2016, February 21). "Clean Air Interstate Rule (CAIR)." Retrieved 3/13/2016, from https://archive.epa.gov/airmarkets/programs/cair/web/html/index.html.

United States Environmental Protection Agency (2016, February 29). "Cross-State Air Pollution Rule (CSAPR)." Retrieved 2/28/2016, from https://www3.epa.gov/crossstaterule/.

United States Environmental Protection Agency (2016, February 22). "Heavy-Duty Highway Diesel Program." Retrieved 4/12/2016, from https://www3.epa.gov/otaq/highway-diesel/regs.htm.

United States Environmental Protection Agency (2016, February 23). "Mercury and Air Toxics Standards (MATS): Cleaner Power Plants." Retrieved 4/12/2016, from https://www3.epa.gov/mats/powerplants.html.

United States Environmental Protection Agency (2016, February 23). "National Trends in Particulate Matter Levels." Retrieved 3/18/2016, from https://www3.epa.gov/airtrends/pm.html.

Veraverbeke, S. (2015). Early season lightning storms followed by vapor pressure deficit anomalies contributed to an extreme wildfire season near the high latitude treeline in Northwest Canada in 2014. 2015 AGU Fall Meeting, American Geophysical Union.

Weaver, B. and K. L. Wuensch (2013). "SPSS and SAS programs for comparing Pearson correlations and OLS regression coefficients." Behavior research methods **45**(3): 880-895.

Zimmerman, D., et al. (1996). "Using trend surface methodology to compare spatial surfaces." United States Department of Agriculture, Forest Service General Technical Report: 129-136.

CHAPTER 4: ASSOCIATION OF THE FRACKING INDUSTRY WITH SMALL SCALE VARIABILITY IN $PM_{2.5}$ POLLUTION IN PENNSYLVANIA, 2004 - 2014

Abstract

Technological advances in directional drilling and fracking have led to a dramatic increase in natural gas production from Pennsylvania (PA). We used geostatistical methods to investigate whether the advent of fracking in PA affected the small scale variation of fine particulate matter ($PM_{2.5}$) pollution across the state. We compared the "pre-fracking" airscape (2004) to the airscape after fracking had been established (2011 and 2014), and failed to identify a significant impact of the fracking industry on mean $PM_{2.5}$ concentration trends in PA from 2004 through 2014. However, the sparse monitoring of $PM_{2.5}$ hindered our ability to detect significant $PM_{2.5}$ trends, particularly in the northeastern region of PA. A sound conclusion regarding the environmental health consequences of fracking in PA must entail additional air quality monitoring as well as an exhaustive investigation of multiple potential pollutants via numerous mediums. The research presented in this chapter is a contribution to this larger investigation. The methods we employed may be used to investigate additional air pollutants associated with fracking industry, including methane and VOC's, to further understand the environmental health risks imposed by fracking.

Abbreviations

AQS	Air quality system
DEP	Pennsylvania Department of Environmental Protection
EPA	United States Environmental Protection Agency
GIS	Geographic information system
PA	Pennsylvania
$PM_{2.5}$	Fine particulate matter, < 2.5 μm in aerodynamic diameter

4.1 Introduction

The decomposition of ancient organic matter produces fossil fuels including natural gas deep under the Earth's surface (Rinaldi 2015). This natural gas is trapped in tiny pores amid large underground shale rock formations. The process of retrieving natural gas from shale and other unconventional gas sources requires the use of hydraulic fracturing (fracking) (Brown, Hartman et al. 2013). Fracking involves injecting high-pressure liquid into the ground to fracture the rock and release the trapped gas. The liquid used in fracking (slickwater) travels deep into the Earth and returns as wastewater that contains total suspended particles (TSP), high salt content, fracturing chemicals, heavy metals, bacteria, and radioactive material, in addition to oil and gas (Brown, Hartman et al. 2013, Easton 2013). The environmental and human health effects of these drilling methods are a topic of much debate (Rahm, Bates et al. 2013, Tillett 2013, De Gouw, Parrish et al. 2014, Rinaldi 2015).

Technological advances in directional drilling and fracking have led to a dramatic increase in natural gas production from unconventional sources (shale, tight sands, and coal bed methane) in the United States (US) (Brown, Hartman et al. 2013, Rahm, Bates et al. 2013, De Gouw, Parrish et al. 2014). Traditionally, coal provided fuel for the US power plants; in 1997, coal comprised over 83% of output from electrical power plant generators (De Gouw, Parrish et al. 2014). But in the last two decades, natural gas has been increasingly used in the electrical power generation sector: in 2014, natural gas contributed 27% of total electrical output,

while coal dropped to 39% (De Gouw, Parrish et al. 2014, U.S. Energy Information Administration 2015).

Pennsylvania (PA) has become a leading source of natural gas in the United States due to the abundance of natural gas from the Marcellus shale, the largest source of shale gas identified in the country (Figure 1) (Brown, Hartman et al. 2013, Rahm, Bates et al. 2013, Muehlenbachs, Spiller et al. 2015, Rinaldi 2015). The number of active wells has increased regularly since 2005 when the first permit for fracking (unconventional drilling) was issued, and PA's contribution to the US natural gas market has flourished with the increased production: in 2013, 13% of the natural gas consumption in the US originated in PA (Brown, Hartman et al. 2013, Muehlenbachs, Spiller et al. 2015, Rinaldi 2015, U.S. Energy Information Administration 2015).

The acceleration of PA's fracking industry continues even as the debate surrounding the environmental health risks of the practice escalates. Scientific evidence supports seemingly opposing conclusions about the risks of fracking, contributing to the murkiness of the controversy. For example, a recent study on the health effects associated with fracking in PA concluded that health symptoms including skin conditions and upper respiratory symptoms increase with increasing proximity to gas wells (Rabinowitz, Slizovskiy et al. 2015), while a recent review article found that similar epidemiological studies "generally lack methodological rigour" (Werner, Vink et al. 2015). In another example, the Environmental Protection Agency (EPA) found few instances of fracking impacts on drinking water and no evidence of systemic, widespread contamination of US drinking water due to

fracking in a draft report released in June 2015 (United States Environmental Protection Agency 2015), while the same year saw New York state ban fracking, citing water resources as a top concern (Webb 2015). Studies have linked fracking with air pollution including methane, volatile organic compounds (VOCs), hazardous air pollutants (HAPs), and ozone, but a lack of air monitors in areas with developed fracking industry hinders effective air quality assessment (Davis 2012, Tollefson 2012, Carlton, Little et al. 2014, Rawlins 2014). Werner et al. (2015) concludes that the gap in the scientific knowledge about the environmental health impacts of fracking "requires urgent attention".

This research contributes to the literature about the environmental health impacts of fracking by investigating the small scale variation in fine particulate matter ($PM_{2.5}$) pollution associated with fracking trends in PA. The health effects of exposure to $PM_{2.5}$ pollution are well established, including increased risk of respiratory illness, aggravation of COPD, bronchitis, asthma, chest pain, and premature mortality (Dockery, Speizer et al. 1989, Pope III, Thun et al. 1995, Laden, Neas et al. 2000, Peng, Bell et al. 2009). We use geostatistical methods to investigate whether the advent of fracking in PA affected the small scale trends of $PM_{2.5}$ pollution within the state. We compare the "pre-fracking" airscape (2004) to the airscape after fracking had been established (2011 and 2014), and hypothesize that different regions within PA experienced different trends in $PM_{2.5}$ pollution during this time period, and that the fracking industry impacted these small scale trends.

4.2 Methods

Daily summary $PM_{2.5}$ concentrations (Code 88101) for each year in the study (2004, 2011, and 2014) were downloaded from the Environmental Protection Agency (EPA) Air Quality System (AQS) website (United States Environmental Protection Agency 2015). We considered the daily $PM_{2.5}$ data in $\mu g/m^3$ for all monitors located in the study area (PA). For each monitor in the study, we averaged the daily data into a monthly average $PM_{2.5}$ value for that monitor.

Fracking well data was downloaded from the PA Department of Environmental Protection (DEP) oil and gas reporting website, accessible at www.paoilandgasreporting.state.pa.us/publicreports. We considered two datasets from DEP, the number of active wells per year and the number of new wells drilled (spudded) per year. Each dataset was limited to unconventional (fracking) wells. The datasets were pulled from two DEP reports: the Wells Drilled by County Report provided the new wells data by year and the Oil and Gas Production Report provided data about active wells by year. Active wells are defined as wells in which a "permit has been issued and well may or may not have been drilled or producing, but has not been plugged" (Pennsylvania Department of Environmental Protection). Unique active wells were identified by unique permit numbers.

We investigated the association between fracking activity and $PM_{2.5}$ pollution via two paired analyses. The first analysis considered the small scale variation in the difference of $PM_{2.5}$ from 2004, before fracking began in PA, and the present (2014), after the establishment of fracking, to investigate how the fracking industry has influenced the $PM_{2.5}$ airscape across PA. The second analysis considered the small

scale variation in the difference from 2004 (before fracking) to 2011, which witnessed the highest number of new unconventional wells established in PA during the study period (Table 1). The process of building a new well, called "spudding", poses a potential for increased particulate pollution. Spudding starts with well construction, which entails an influx of industrial vehicles and machinery into an area and drilling deep into the Earth to access the Marcellus shale (Rinaldi 2015). Once the well is constructed, it is "stimulated": explosives are detonated in the well bore and fracking fluid is pumped through the well at high pressure to release the natural gas from the shale rock; the fracking fluid returns to the Earth's surface carrying suspended particles and other contaminants (Brown, Hartman et al. 2013, Easton 2013, Rinaldi 2015). The potential for $PM_{2.5}$ pollution may arise at any or all of these steps in spudding a new well.

The spatial surface comparison methods for both the current (2004 vs. 2014) and spudding (2004 vs. 2011) analyses followed those detailed in Chapter 3 of this dissertation document. The geographic information system QGIS (version 2.10.1-Pisa) was used to generate a prediction grid of 4,070 evenly distributed points over the study area, which covered all 67 counties in PA. We considered the summer month data (June, July, and August) for 2004 (before fracking) and 2014 (current analysis) or 2011 (spudding analysis). Ordinary kriging was used to predict $PM_{2.5}$ concentrations at "missing" monitor locations for each year, so that a paired geostatistical dataset with $PM_{2.5}$ concentrations at coincident monitor locations was created for both the before (2004) and after (2014 or 2011) fracking datasets. The difference in $PM_{2.5}$ concentrations was then calculated at each monitor location.

Ordinary kriging was performed to extend these observations into predicted differences in $PM_{2.5}$ concentrations at every point in the prediction grid and then smoothed to create a predicted difference surface across PA. To account for *edge effects* – the influence of events close to but outside of the study region (Schabenberger and Gotway 2005) – we included data from monitors located within 100 km of the PA state border to enhance the kriged predictive power in the border regions. The significance of the predicted differences was assessed using the conditional simulation approach developed in Chapter 3. Briefly, 1,000 mapped realizations (multivariate simulations from the statistical conditional kriged predictive distributions) were generated at each prediction grid location. This distribution of prediction realizations of the $PM_{2.5}$ concentration differences was used to estimate the probability that the predicted surface was different from zero, thus signifying a change in $PM_{2.5}$ concentrations from before (2004) to after (2014 / 2011) fracking. This probability can be interpreted as a p-value and used to judge statistical significance by comparing its complement (one minus the probability) to fixed alpha or type I error rates. We considered the error rate $\alpha < 0.05$ to be ample evidence of a significant difference in $PM_{2.5}$ concentrations from 2004 to 2014 / 2011. We generated maps showing the predicted differences in $PM_{2.5}$ concentrations, and maps of the significance levels of these predictions and of the predicted standard errors. Aligning these maps with the location data of current and new fracking activities allowed us to visualize associations between small scale variations in $PM_{2.5}$ concentrations and fracking across PA.

4.3 Results

Table 1 reports summary data for PA by year. The yearly mean $PM_{2.5}$ across the state of PA generally decreased from year to year, from 14.25 $\mu g/m^3$ in 2004 to 11.57 $\mu g/m^3$ in 2011 to 10.67 $\mu g/m^3$ in 2014 (Table 1). This follows the trends previously reported for the NE US in Chapter 3 of this dissertation document and in previous research (Saunders and Waugh 2015). The maximum $PM_{2.5}$ concentration values show more variability year to year, with a high maximum $PM_{2.5}$ concentration of 30.93 $\mu g/m^3$ in 2005 and a low maximum of 18.43 $\mu g/m^3$ in 2012 (Table 1). This value is heavily influenced by single high $PM_{2.5}$ events, such as wild fires, and therefore the variability is expected (Saunders and Waugh 2015).

In June 2004, 100 monitors reported $PM_{2.5}$ concentrations in the study area. In June 2014, 94 monitors reported $PM_{2.5}$ concentrations, while in 2011, 96 monitors reported observations. The locations of these monitors vary from year to year, yielding incomplete paired spatial designs. Ordinary kriging was used to predict $PM_{2.5}$ concentrations at the 25 "missing" monitor locations for June 2004 and at the 31 "missing" monitor locations for June 2014 in the current analysis. In the spudding analysis, ordinary kriging predicted $PM_{2.5}$ concentrations at the 21 "missing" monitor locations for June 2004 and at the 25 "missing" monitor locations for June 2011. Thus, a paired geostatistical dataset with $PM_{2.5}$ concentrations at coincident monitor locations was created for both the current (2004 vs. 2014) and spudding (2004 vs. 2011) analyses for June, with N= 125 observations in the current analysis and N = 121 observations in the spudding analysis (Table 2). The average difference in $PM_{2.5}$ concentrations from June 2004 to June 2014 was -5.42 $\mu g/m^3$

(range = -11.36 to 1.79), with 98.41% of observation locations showing a decrease in $PM_{2.5}$ (Table 2). The average difference in $PM_{2.5}$ concentrations from June 2004 to June 2011 was -1.98 μg/m^3 (range = -8.83 to 7.64), with 74.38% of observation locations showing a decrease in $PM_{2.5}$ (Table 2).

In July 2004, 100 monitors reported $PM_{2.5}$ concentrations in the study area. In July 2014, 96 monitors reported $PM_{2.5}$ concentrations, while in 2011, 97 monitors reported observations. Ordinary kriging was used to predict $PM_{2.5}$ concentrations at the 26 "missing" monitor locations for July 2004 and at the 30 "missing" monitor locations for July 2014 in the current analysis. In the spudding analysis, ordinary kriging predicted $PM_{2.5}$ concentrations at the 21 "missing" monitor locations for July 2004 and at the 24 "missing" monitor locations for July 2011. The paired geostatistical dataset with $PM_{2.5}$ concentrations at coincident monitor locations for July included N= 126 observations in the current analysis and N = 121 observations in the spudding analysis (Table 2). The average difference in $PM_{2.5}$ concentrations from July 2004 to July 2014 was -10.62 μg/m^3 (range = -17.71 to -3.00), with 100% of locations showing a decrease in $PM_{2.5}$ (Table 2). The average difference in $PM_{2.5}$ concentrations from July 2004 to July 2011 was -4.97 μg/m^3 (range = -12.56 to 3.90), with 95.87% of locations showing a decrease in $PM_{2.5}$ (Table 2).

In August 2004, 101 monitors reported $PM_{2.5}$ concentrations in the study area. In August 2014, 96 monitors reported $PM_{2.5}$ concentrations, while in 2011, 97 monitors reported observations. Ordinary kriging was used to predict $PM_{2.5}$ concentrations at the 25 "missing" monitor locations for August 2004 and at the 30 "missing" monitor locations for August 2014 in the current analysis. In the

spudding analysis, ordinary kriging predicted $PM_{2.5}$ concentrations at the 20 "missing" monitor locations for August 2004 and at the 24 "missing" monitor locations for August 2011. The paired geostatistical dataset with $PM_{2.5}$ concentrations at coincident monitor locations for August was comprised of N= 126 observations in the current analysis and N = 121 observations in the spudding analysis (Table 2). The average difference in $PM_{2.5}$ concentrations from August 2004 to August 2014 was -7.42 $\mu g/m^3$ (range = -16.87 to -1.14), with 100% of observations showing a decrease in $PM_{2.5}$ (Table 2). The average difference in $PM_{2.5}$ concentrations from August 2004 to August 2011 was -5.83 $\mu g/m^3$ (range = -12.66 to -0.16), with 100% of observations showing a decrease (Table 2).

Only the July outcomes of the current and spudding small scale variation analyses are presented in the remainder of this document; results and interpretations from the small scale variation analyses for June and August were similar despite the differences in the means and ranges noted above. Figure 2 shows the small scale variation analysis comparing $PM_{2.5}$ pollution differences from 2004 (pre-fracking) to the current (2014) airscape. The predicted surface suggests a decrease in $PM_{2.5}$ pollution concentrations from 2004 to 2014 across all regions within PA (average predicted decrease = -8.43 $\mu g/m^3$). The largest decreases were located in the south central area of the state. Much of this area overlays the Marcellus shale, and there are active fracking wells located in this area of most improved $PM_{2.5}$ pollution. The northern regions of PA experienced the least improvement in $PM_{2.5}$ pollution concentrations from 2004 – 2014. The Marcellus shale underlies the entirety of this least-improved area, and there is a cluster of

active fracking wells located in the northeastern region. Figure 3 reveals that all of the predicted differences in $PM_{2.5}$ concentrations from 2004 to 2014 were significant; the least significant differences align with the lowest difference in predicted $PM_{2.5}$ (the northern most region of the state, Figures 2 and 3). Figure 4 presents the standard errors (SE's) of the predicted differences in $PM_{2.5}$ concentrations from 2004 to 2014. As expected, the SE's vary across the state in accordance with the spatial distribution of the $PM_{2.5}$ monitors. Areas with few or no $PM_{2.5}$ monitors show the highest SE's of the predicted differences in $PM_{2.5}$. The highest SE's are seen in the far northeastern and the far western regions of PA, far removed from the $PM_{2.5}$ monitors in the south and southeast areas of the state. Notably, both areas rest atop the Marcellus shale (Figure 1), and the fracking industry is well established in the northeastern region (Figure 4).

Figure 5 exhibits the small scale variation analysis comparing $PM_{2.5}$ pollution differences from 2004 (pre-fracking) to the highest spudding year (2011). 1,619 new fracking wells were constructed in 2011 (Table 1). The predicted surface suggests a decrease in $PM_{2.5}$ pollution concentrations from 2004 to 2011 across all regions within PA (average predicted decrease = 3.52 $\mu g/m^3$); however, the northern regions showed only a slight predicted decrease of 1.00 $\mu g/m^3$ and much of this area failed to achieve significant decreases (Figure 5). This area of low improvement includes counties with new fracking wells spudded in 2011; contrarily, almost half of the counties with new well spudding are located in regions that showed average or greater improvement in $PM_{2.5}$ pollution from 2004 to 2011. Figure 6 shows the southern half of the state experienced significant predicted

decreases in $PM_{2.5}$ pollution concentrations from 2004 to 2011, including half of the counties with new well spudding. Figure 7 indicates that the standard errors (SE's) of the predicted differences in $PM_{2.5}$ concentrations from 2004 to 2011 follow the spatial distribution of the $PM_{2.5}$ monitors, with the lowest SE's in the more densely monitored southern regions and the highest SE's in the northeastern and the far western regions where there are no $PM_{2.5}$ monitors.

4.4 Discussion

The analyses presented in this chapter failed to identify a significant impact of the fracking industry on the small scale variation in $PM_{2.5}$ concentrations in PA from 2004 through 2014. PA experienced the same overall decrease in $PM_{2.5}$ pollution identified across the NE US during this time, and no systematic differences in this trend were discovered comparing areas with to areas without active fracking wells in PA. Surprisingly, an investigation into a year with considerable new well activity (2011) also failed to detect a significant association between areas with increased new well construction (spudding) and changes in small scale $PM_{2.5}$ variation trends. However, in both the current (2014) and the spudding (2011) analyses, the sparse monitoring of $PM_{2.5}$ hindered our ability to detect associations between fracking and $PM_{2.5}$ trends, particularly in the northeastern region of PA, where fracking industry is well established but $PM_{2.5}$ monitors are all but absent (Figures 2 and 5). Therefore, additional air quality monitoring focused on areas with current and future fracking production is necessary to gain a more complete understanding of the air quality risks associated with fracking in PA.

The northeastern region of PA showed the lowest predicted differences in $PM_{2.5}$ during our study period, and much of this region failed to identify a significant difference from 2004 to 2011 (Figures 2 and 5). This area remained the least improved and one of the least monitored regions in PA in 2014. Due to its sparse monitoring for $PM_{2.5}$, the northeast region joins the far western region in displaying the largest estimates of prediction uncertainty in the study area (Figures 4 and 7). Considering its location atop the Marcellus shale and the clustering of active fracking wells in the northeast (Figure 2), a more robust network of $PM_{2.5}$ and other air quality monitors in this area is imperative to accurately capture the air pollution in this fracking-heavy region.

A recent public opinion poll conducted by the University of Michigan found that PA residents felt uncertain about the environmental and health risks associated with fracking (Brown, Hartman et al. 2013). The lack of significant evidence of an association between fracking and $PM_{2.5}$ pollution identified in this research may bring some comfort to PA residents and other invested parties; however, the limitations of these results, particularly the need for increased $PM_{2.5}$ monitors to reduce prediction uncertainty in the fracking-heavy northeastern region of PA, should be noted. A sound conclusion regarding the environmental health consequences of fracking in PA must entail an exhaustive investigation of multiple potential pollutants via numerous mediums, including air, soil, and water. The inclusion of additional air quality monitors in areas with active fracking is necessary to continue the research presented in this chapter. Thus, this research is a prelude

for the evidence required to reach a substantiated conclusion about the environmental health risks imposed by fracking.

Limitations

As in the previous chapters of this dissertation document, the research presented in this chapter relies on previously collected data and utilizes existing data to estimate values for missing measurements. Thus, the conclusions are limited by the methods and reporting of the primary data collections. Notably, the well production data reported by DEP is based on self-reports by the gas companies operating in PA (Pennsylvania Department of Environmental Protection). Appendix C contains further discussion of the DEP data limitations.

This research utilized the daily summary $PM_{2.5}$ concentrations from EPA's AQS and averaged the daily data into a monthly average $PM_{2.5}$ value for each monitor in the study. We failed to conclude that the presence of fracking industry significantly impacts mean $PM_{2.5}$ concentration trends in PA from 2004 through 2014; however, an investigation into peak $PM_{2.5}$ concentrations recorded by monitors in the AQS may uncover an association between fracking events and local, short term spikes in $PM_{2.5}$ pollution.

It should be noted that while the $PM_{2.5}$ pollution predictions follow a geostatistical framework described above to ascertain a measure of prediction uncertainty, the association between fracking locations and the predicted difference in $PM_{2.5}$ concentrations are based on visual inspections of the mapped data (Figures 2 – 7).

Strengths

The primary strength of the research presented in this chapter is the utilization of a paired geospatial methodology to investigate the small scale air pollution changes associated with fracking. We considered particulate pollution in this research; previous research has identified other air pollutants associated with fracking and other natural gas production activities, including methane and volatile organic compounds (VOCs) (Howarth, Santoro et al. 2011, De Gouw, Parrish et al. 2014). The association of methane and VOC pollution with natural gas production is gaining attention: on August 18, 2015, the EPA proposed measures to reduce methane and VOC emissions from the natural gas industry, and at the time of this writing, EPA has begun the Information Collection Request (ICR) process, requiring gas companies to provide information that will be used in the development of regulations to reduce methane and VOC emissions from natural gas production activities (United States Environmental Protection Agency 2016). The methodologies described in this dissertation chapter may be extended to investigate the association between fracking and these air pollutants in PA and in other states with active fracking industry to investigate the impacts of these imminent regulations.

4.5 Conclusion

We analyzed the small scale variation in $PM_{2.5}$ pollution in PA to compare the "pre-fracking" airscape (2004) to the airscape after fracking had been established (2011 and 2014) to test our hypotheses that different regions within PA

experienced different trends in $PM_{2.5}$ pollution during this time period and that the fracking industry impacted these trends. We failed to conclude that either the presence of fracking industry or a significant increase in new well construction in a single year significantly affected the mean $PM_{2.5}$ concentration trends in PA from 2004 through 2014. However, the sparse monitoring of $PM_{2.5}$ concentrations hindered our ability to detect associations between fracking and $PM_{2.5}$ trends, particularly in the northeastern region of PA. Additional air quality monitoring focused on areas with current and future fracking production is necessary to gain a more complete understanding of the air quality risks associated with fracking in PA.

4.6 Tables and Figures

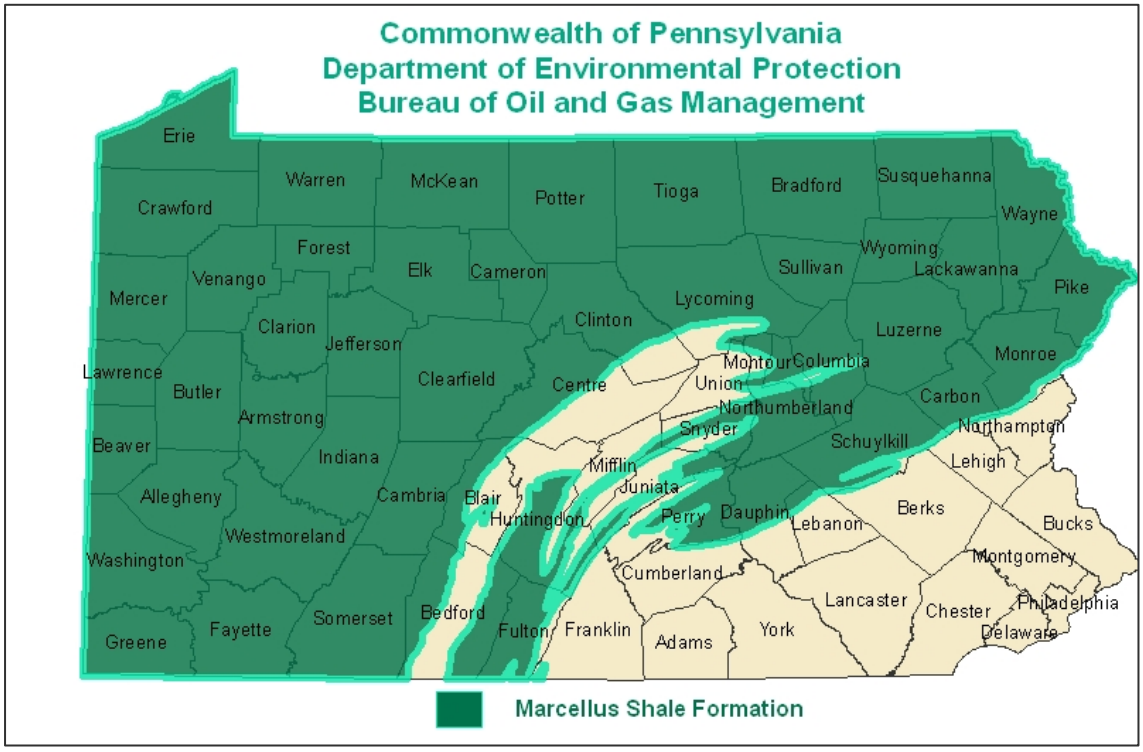

Figure 1. Image of Marcellus shale boundaries in the state of Pennsylvania. Reprinted with permission from the Pennsylvania Department of Environmental Protection, Office of Oil and Gas Management, Bureau of Compliance and Data Administration[1].

[1]Email communication with Myron Suchodolski, Environmental Program Manager, and Roger Dietz, Information Technology Generalist, 4/25 and 4/26/16

Year	Mean PM2.5	Minimum PM2.5	Maximum PM2.5	Difference in Mean	Difference in Maximum	# New Unconventional Wells
2004	14.25	6.31	27.53	-0.71	-0.69	0
2005	15.29	6.45	30.93	1.04	3.40	3
2006	13.44	6.20	29.69	-1.85	-1.24	15
2007	14.16	5.80	29.09	-1.30	-0.60	68
2008	12.86	6.50	28.22	-1.30	-0.87	269
2009	11.56	5.96	24.11	-1.30	-4.11	723
2010	12.01	5.80	24.14	0.45	0.03	1267
2011	11.57	3.19	25.04	-0.44	0.90	1619
2012	10.62	5.19	18.43	-0.95	-6.61	1179
2013	10.36	4.41	23.17	-0.26	4.74	1147
2014	10.67	3.70	27.25	0.31	4.08	1281

Table 1. Summary of $PM_{2.5}$ and new unconventional (fracking) wells in PA by year. The difference in mean and maximum values compares the values for the associated year with the values for the prior year.

	Current Analysis (2004 vs. 2014)			Spudding Analysis (2004 vs. 2011)		
	June	July	August	June	July	August
Average (ug/m3)	-5.52	-10.62	-7.42	-1.98	-4.97	-5.83
Minimum (ug/m3)	-11.36	-17.71	-16.87	-8.83	-12.56	-12.66
Maximum (ug/m3)	1.79	-3.00	-1.14	7.64	3.90	-0.16
# Observations	125	126	126	121	121	121
% Observations < 0	98.41%	100.00%	100.00%	74.38%	95.87%	100.00%

Table 2. Summary of differences in $PM_{2.5}$ concentrations at coincident monitor locations for the current and spudding analyses of the summer months. The current analysis subtracts the 2004 (before fracking) $PM_{2.5}$ values from the current (2014) values. The spudding analysis subtracts the 2004 $PM_{2.5}$ values from the 2011 values. Percent observations less than 0 indicates the number of monitors that recorded an improvement in $PM_{2.5}$, with lower values at the end of the analysis observation period (2014 or 2011) compared to values at the beginning of the analysis observation period (2004).

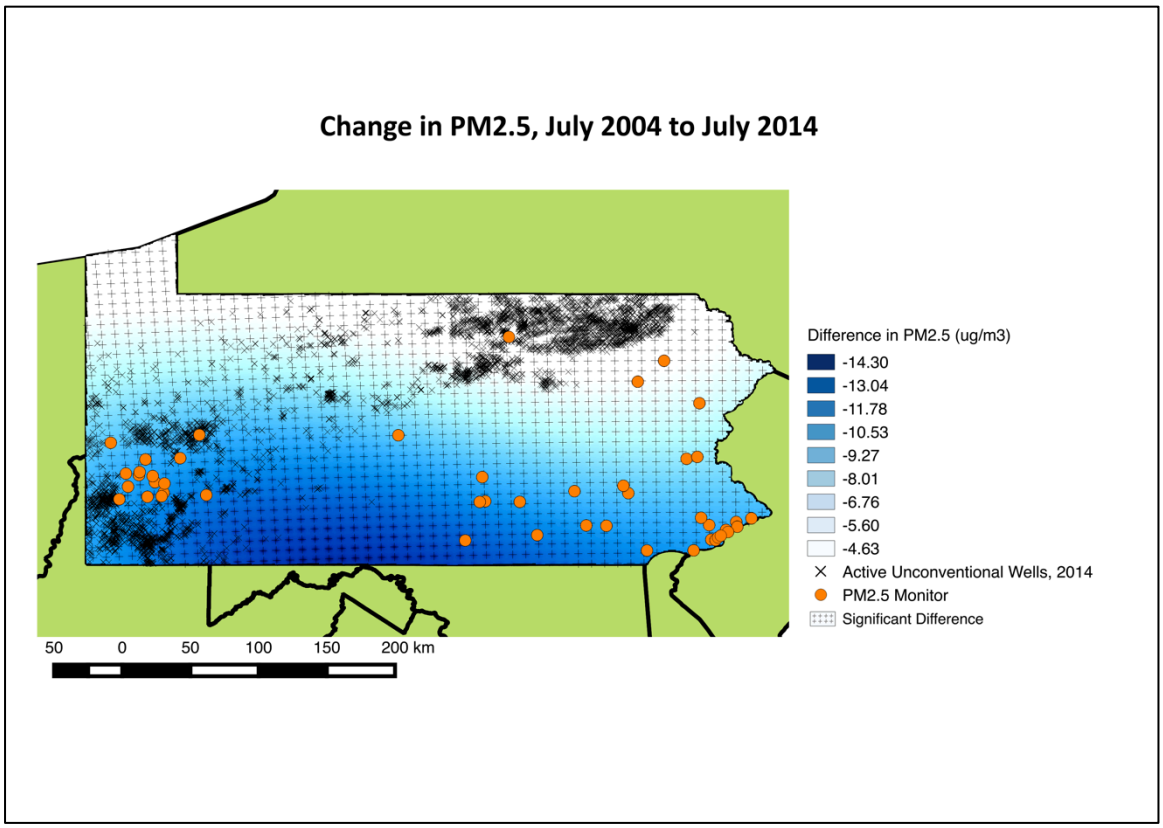

Figure 2. Map of the ordinary kriged predicted differences in $PM_{2.5}$ concentrations from July 2004 to July 2014. Significance is determined at $\alpha = 0.05$.

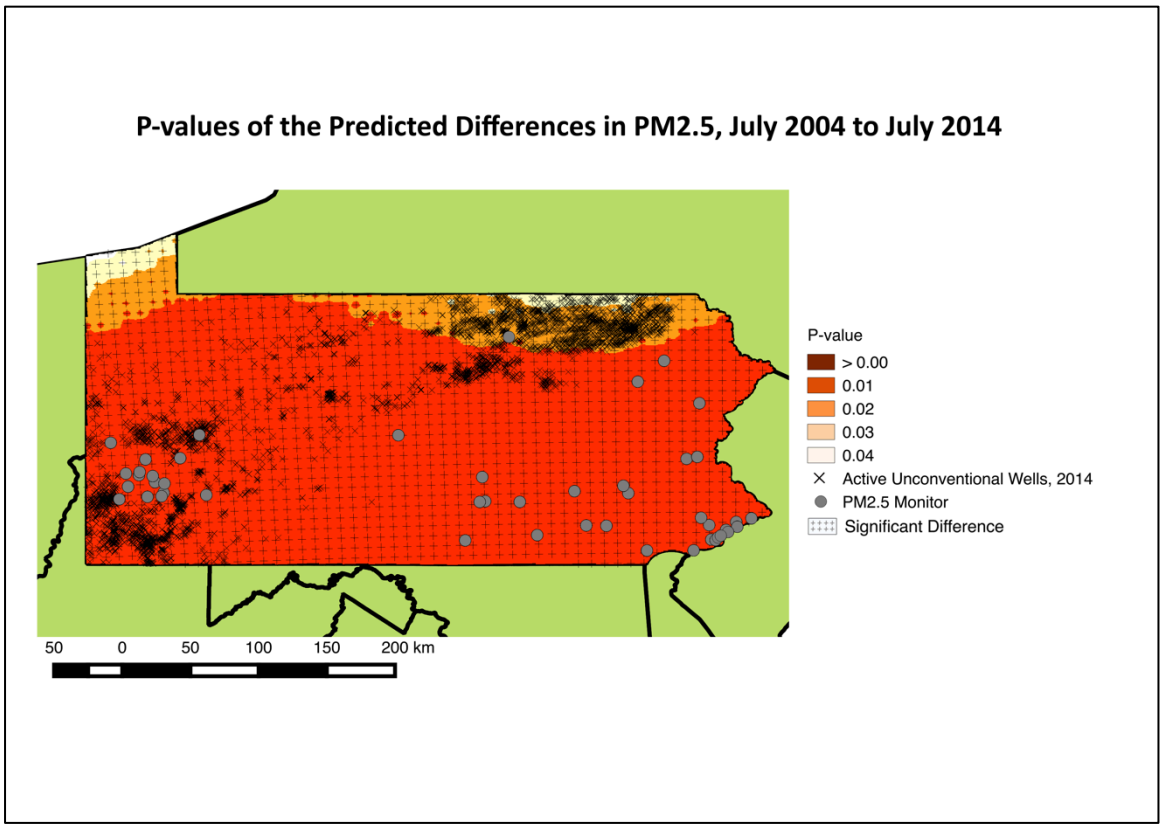

Figure 3. Map of the p-values of the predicted differences in $PM_{2.5}$, July 2004 – July 2014. Significance is determined at $\alpha = 0.05$.

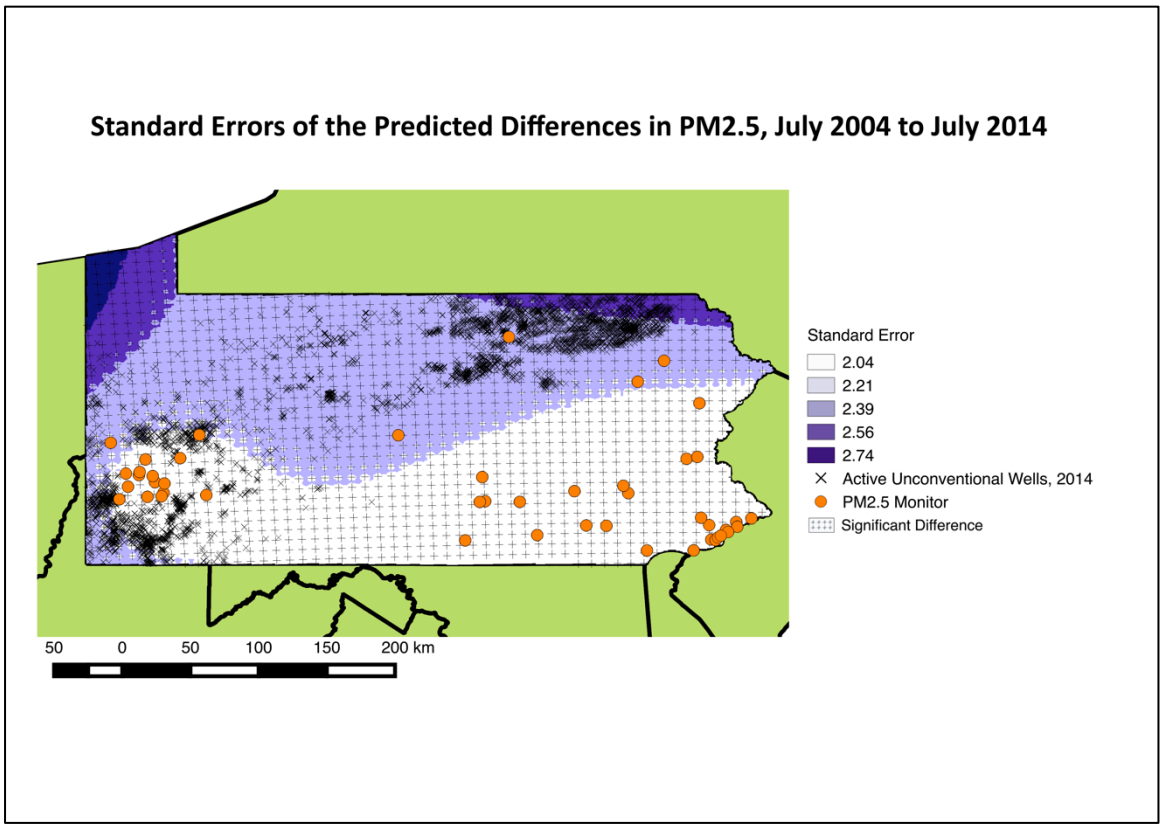

Figure 4. Map of the standard error of the predicted differences in $PM_{2.5}$, July 2004 – July 2014. Significance is determined at $\alpha = 0.05$.

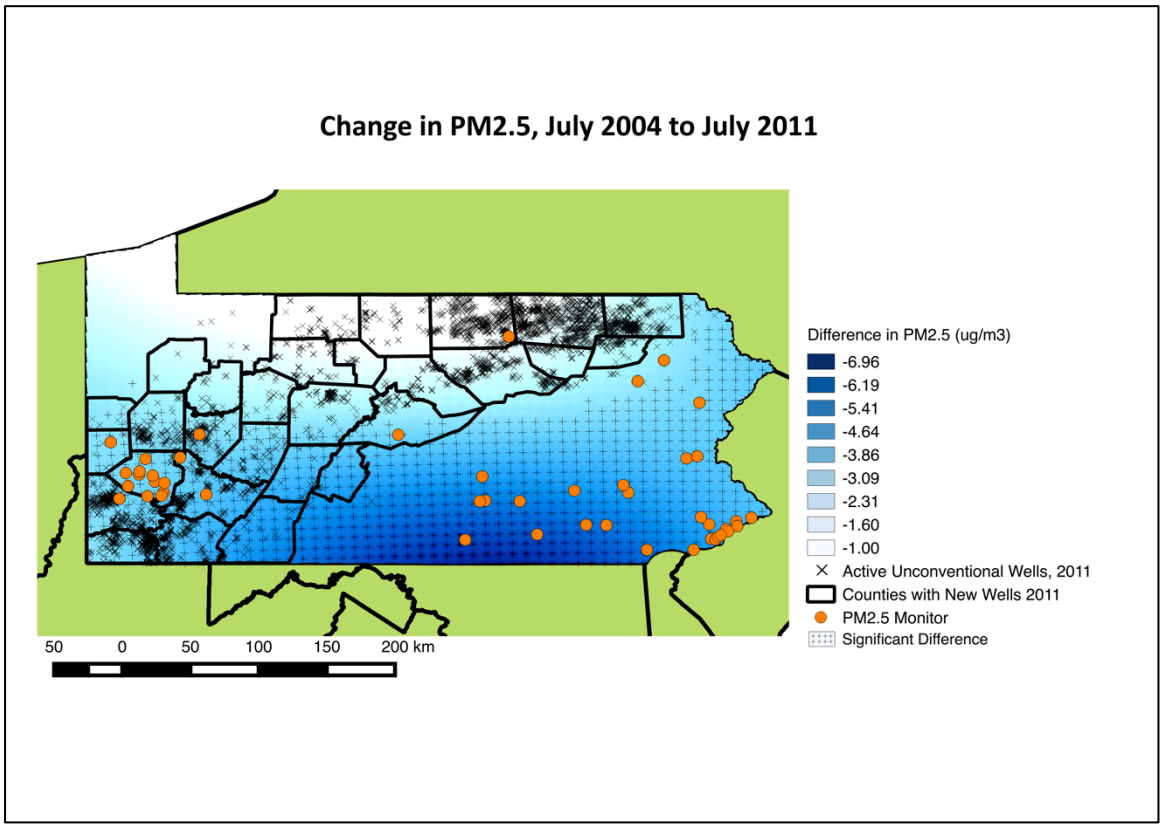

Figure 5. Map of the ordinary kriged predicted differences in $PM_{2.5}$ concentrations from July 2004 to July 2011. Significance is determined at $\alpha = 0.05$. Counties with unconventional gas wells built in 2011 are outlined.

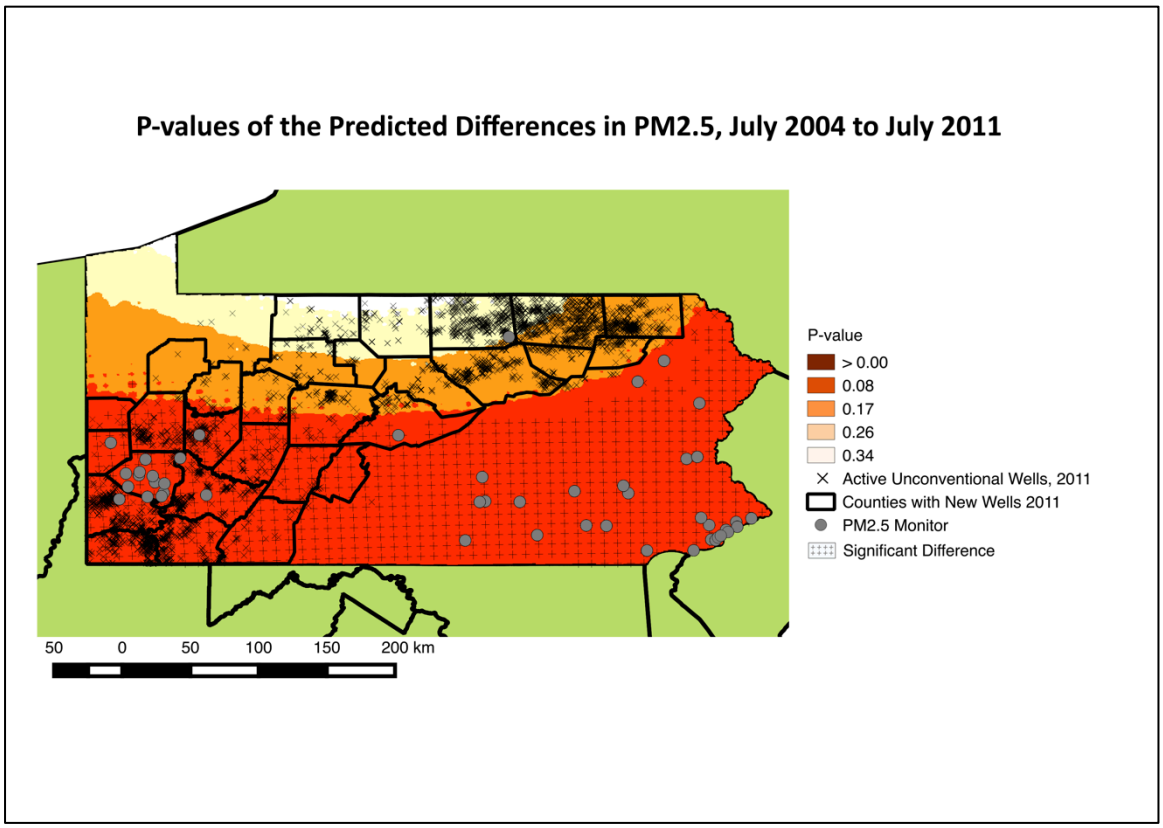

Figure 6. Map of the p-values of the predicted differences in $PM_{2.5}$, July 2004 – July 2011. Significance is determined at $\alpha = 0.05$. Counties with unconventional gas wells built in 2011 are outlined.

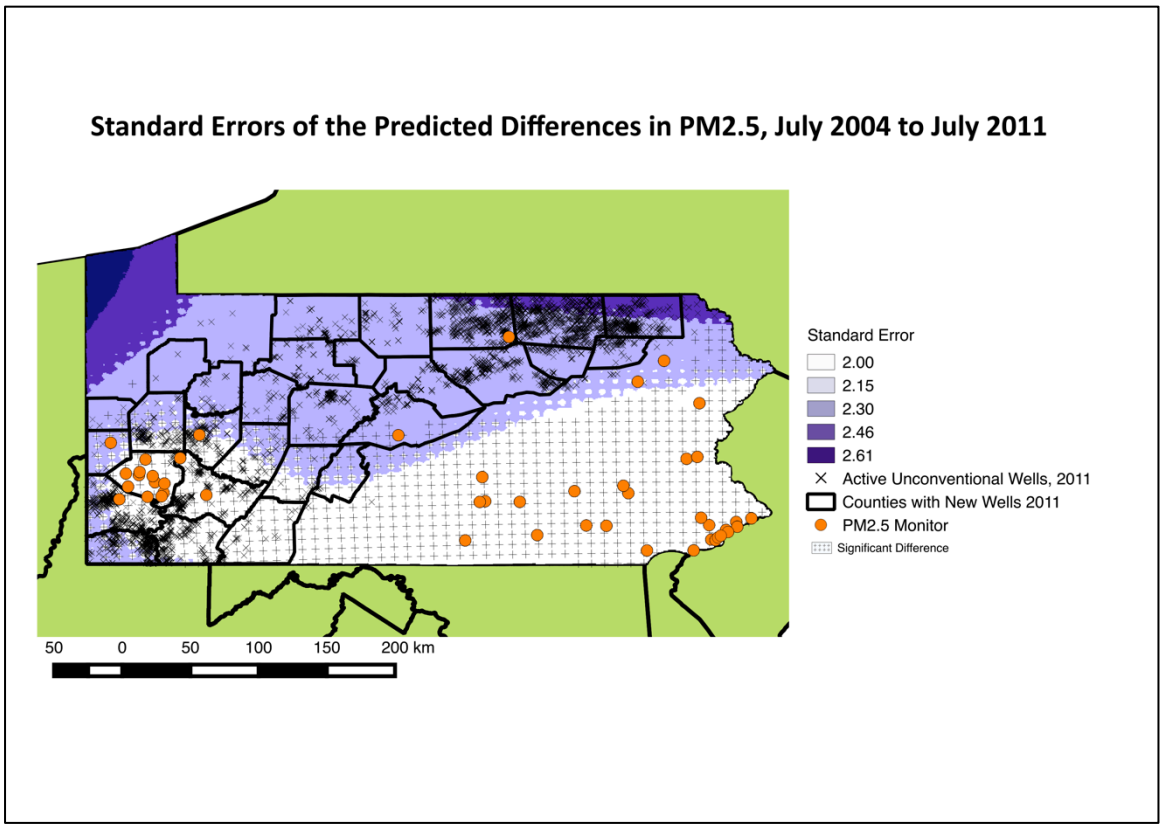

Figure 7. Map of the standard error of the predicted differences in $PM_{2.5}$, July 2004 – July 2011. Significance is determined at $\alpha = 0.05$. Counties with unconventional gas wells built in 2011 are outlined.

4.7 References

Brown, E., et al. (2013). "The national surveys on energy and environment public opinion on fracking: Perspectives from Michigan and Pennsylvania." Center for Local, State, and Urban Policy (CLOSUP),'Survery Report: Climate Policy Options.

Carlton, A. G., et al. (2014). "The data gap: can a lack of monitors obscure loss of clean air act benefits in fracking areas?" Environmental science & technology **48**(2): 893-894.

Davis, C. (2012). "The politics of "fracking": Regulating natural gas drilling practices in Colorado and Texas." Review of Policy Research **29**(2): 177-191.

De Gouw, J., et al. (2014). "Reduced emissions of CO2, NOx, and SO2 from US power plants owing to switch from coal to natural gas with combined cycle technology." Earth's Future **2**(2): 75-82.

Dockery, D. W., et al. (1989). "Effects of inhalable particles on respiratory health of children." American Review of Respiratory Disease **139**(3): 587-594.

Easton, J. (2013). "Fracking Wastewater Management." Water & Wastewater International **28**(5).

Howarth, R. W., et al. (2011). "Methane and the greenhouse-gas footprint of natural gas from shale formations." Climatic Change **106**(4): 679-690.

Laden, F., et al. (2000). "Association of fine particulate matter from different sources with daily mortality in six US cities." Environmental Health Perspectives **108**(10): 941.

Muehlenbachs, L., et al. (2015). "The Impact of the Fracking Boom on Rents in Pennsylvania Working Paper." Retrieved 4/14/2016, from http://public.econ.duke.edu/~timmins/fracking_rents.pdf.

Peng, R. D., et al. (2009). "Emergency admissions for cardiovascular and respiratory diseases and the chemical composition of fine particle air pollution." Environmental Health Perspectives **117**(6): 957.

Pennsylvania Department of Environmental Protection. Oil and Gas Reports Data Dictionary.

Pope III, C. A., et al. (1995). "Particulate air pollution as a predictor of mortality in a prospective study of US adults." American journal of respiratory and critical care medicine **151**(3_pt_1): 669-674.

Rabinowitz, P. M., et al. (2015). "Proximity to natural gas wells and reported health status: Results of a household survey in Washington County, Pennsylvania." Environmental Health Perspectives (Online) **123**(1): 21.

Rahm, B. G., et al. (2013). "Wastewater management and Marcellus Shale gas development: trends, drivers, and planning implications." Journal of environmental management **120**: 105-113.

Rawlins, R. A. (2014). "Planning for fracking on the Barnett shale: Urban air pollution, improving health based regulation, and the role of local governments." Retrieved 4/14/2016, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2410648.

Rinaldi, R. (2015). "Fracturing the Keystone: Why Fracking in Pennsylvania Should Be Considered an Abnormally Dangerous Activity." Widener Law Journal **24**(2).

Saunders, R. O. and D. W. Waugh (2015). "Variability and potential sources of summer PM 2.5 in the Northeastern United States." Atmospheric Environment **117**: 259-270.

Schabenberger, O. and C. A. Gotway (2005). Statistical methods for spatial data analysis, CRC press.

Tillett, T. (2013). "Summit discusses public health implications of fracking." Environmental Health Perspectives **121**(1): a15.

Tollefson, J. (2012). "Air sampling reveals high emissions from gas field." Nature **482**(7384): 139-140.

U.S. Energy Information Administration (2015). December 2015 Monthly Energy Review. Monthly Energy Review. B. T. Fichman. http://www.eia.gov/totalenergy/data/monthly/.

U.S. Energy Information Administration (2015). " Where Our Natural Gas Comes From." Retrieved June 1, 2016, 2016, from http://www.eia.gov/energyexplained/index.cfm?page=natural_gas_where.

United States Environmental Protection Agency (2015, December 4). "Air Data." Retrieved 12/7/2015, from https://www3.epa.gov/airquality/airdata/.

United States Environmental Protection Agency (2015). DRAFT: Hydraulic Fracturing Drinking Water Assessment Executive Summary.

United States Environmental Protection Agency (2016, April 14). "Methane : Addressing Greenhouse Gases and Smog forming VOCs from the Oil and Gas

Industry." Retrieved 4/29/2016, from https://www3.epa.gov/airquality/oilandgas/methane.html.

Webb, R. (2015). "Better Safe Than Sorry: New York Bans Fracking Due to Potential Impacts on Water Resources."

Werner, A. K., et al. (2015). "Environmental health impacts of unconventional natural gas development: A review of the current strength of evidence." Science of The Total Environment **505**: 1127-1141.

CHAPTER 5: CONCLUSIONS

5.1 Dissertation Overview

This dissertation characterized the variation in $PM_{2.5}$ over space and time in the northeastern United States (NE US) from 2000 to 2014. Specific research aims were proposed to accomplish this objective, and each research aim was presented in a separate chapter in the dissertation document.

Chapter 2

Chapter 2 presented dissertation research Aim 1: to characterize the spatial-temporal variation in fine particulate matter ($PM_{2.5}$) in the NE US from 2000-2014 as a function of environmental determinants. To meet this aim, we employed non-spatial and spatial statistical modeling techniques to identify significant environmental determinants associated with $PM_{2.5}$ pollution. Environmental covariates were considered at different spatial aggregations, including monitor and power plant locations at the point-level, population, toxic releases, and elevation at the county level, and energy generation and annual traffic at the state level of spatial aggregation. The temporal covariates of season and year were also explored. We undertook a deliberate, step-wise approach to build our final model, comparing model performance and significance of covariates as well as investigating the model residual spatial dependence, to arrive at a statistically sound working model for $PM_{2.5}$ in the NE US. Our analysis concluded that the environmental determinants of monitor Y coordinate, power plant location, elevation, energy generation, traffic, season, and year adequately explained the spatial variation in $PM_{2.5}$ in the NE US from 2000 – 2014 a multilevel model with a constant variance (error term). The

final model that included these significant components was employed in the subsequent research aim described in Chapter 3 of this dissertation document.

Chapter 3

Chapter 3 encompassed dissertation research Aim 2: to investigate large scale trends and small scale spatial variation in $PM_{2.5}$ pollution and the efficacy of federal emissions regulations in reducing $PM_{2.5}$ pollution in the NE US. We examined U.S. Environmental Protection Agency (EPA) regulations designed to reduce $PM_{2.5}$ and precursor emissions from two critical sources of particulate pollution in the NE US: power plants and mobile sources including passenger and heavy duty vehicles. We investigated how the relationship of these sources to $PM_{2.5}$ changed from 2000 to 2014 and introduced an innovative approach to assess significant small scale changes in $PM_{2.5}$ pollution across the NE US during this period of changing emission regulations. The analysis showed that while the NE US experienced an overall decrease in $PM_{2.5}$ concentrations from 2000 to 2014, regions within the NE US experienced different trends in the $PM_{2.5}$ airscape during this time. Furthermore, we concluded that regulations aimed at power plant emissions significantly decrease $PM_{2.5}$ pollution.

Chapter 4

Chapter 4 illustrated dissertation research Aim 3: to explore whether the establishment of the fracking industry in Pennsylvania (PA) impacted the small scale spatial variability in $PM_{2.5}$ pollution within the state from 2004 to 2014. We used geostatistical methods to compare the "pre-fracking" airscape (2004) to the

airscape after fracking had been established (2011 and 2014). We failed to identify a significant impact of the fracking industry on $PM_{2.5}$ concentration trends in PA from 2004 through 2014; however, the sparse monitoring of $PM_{2.5}$ hindered our ability to detect significant $PM_{2.5}$ trends, particularly in the northeastern region of PA where fracking is well established but $PM_{2.5}$ monitors are absent. We concluded that additional air quality monitors and robust investigations of multiple air, water, and soil pollutants are required to understand the environmental health risks imposed by fracking.

5.2 Strengths and Limitations

The research presented in this dissertation document is secondary data analysis. We used publicly available data from multiple governmental agencies to investigate the space-time variation in $PM_{2.5}$ pollution. The outcome variable ($PM_{2.5}$) used in all chapters as well as the model predictor variables (power plant locations, population, toxic releases, elevation, energy generation, traffic measures) utilized in Chapters 2 and 3 were gathered from federal agencies including EPA, the U.S. Census Bureau, the U.S. Geological Survey (USGS), the U.S. Energy Information Administration (EIA), and the Federal Highway Administration (FHWA). The fracking data analyzed in Chapter 4 originated from the Pennsylvania Department of Environmental Protection (DEP). The conclusions reached by the analyses in this dissertation are therefore limited by the methods and reporting of the primary data collections. Furthermore, data were not available for every year and at every

location in our study. We utilized existing data to estimate missing values as described in Chapters 2, 3, and 4.

In Chapter 3, we considered only the EPA federal regulations that introduced stricter power plant and mobile source standards from 2000 to 2014 and assumed that these standards have the greatest impact on the relationships between the associated environmental determinants and $PM_{2.5}$ pollution in our study area. However, states and localities may enact stricter rules beyond the limits of the EPA power plant and mobile source standards, which could impact the relationship between these environmental determinants and $PM_{2.5}$ concentrations in the NE US. We did not control for the fleet makeup of the yearly traffic indicators; the influence of the traffic variable may differ across time and space as newer vehicle models come into use with updated emission technologies. We also did not assess high emission events such as wildfires that may further influence $PM_{2.5}$ concentrations in the NE US in the study area.

The association between fracking industry and the predicted difference in $PM_{2.5}$ concentrations in PA explored in Chapter 4 are based on visual inspections of the mapped data and do not entail a statistical measure of prediction uncertainty. A suggestion for future research that extends the methodology presented in Chapter 4 into a large scale trend analysis is outlined below (see "Future Directions").

The strengths of this research rest on the deliberate, step-wise process we undertook to build the multivariate model (Chapter 2) and in the innovative methods we introduce in the small scale spatial variation analyses (Chapters 3 and 4). The statistical analyses employed throughout the dissertation recognize the

spatial nature of $PM_{2.5}$ pollution and account for the spatial dependence structure. This approach assured statistically appropriate estimations and interpretations of hypothesis tests and avoided the potential for spurious effects as seen in analyses that fail to account for residual spatial variation (Cressie 1993, Schabenberger and Gotway 2005, Bivand, Pebesma et al. 2008, Berman, Breysse et al. 2015).

5.3 Contributions to Public Health

Air quality falls under the purview of EPA, and the Agency passes legislation to reduce air pollution in accordance with its mission statement: to protect human health and the environment (United States Environmental Protection Agency 2015). The research presented in this dissertation document uses data collected and disseminated by EPA to investigate the air quality of the NE US. We utilized spatial statistical techniques to complete maps of the $PM_{2.5}$ airscape by predicting $PM_{2.5}$ concentrations in areas that lack air quality monitors in the EPA's Air Quality System (AQS) network. Thus, this research contributes to the understanding of the extent of $PM_{2.5}$ pollution in the NE US.

This dissertation research investigates the impact of EPA laws and regulations that seek to reduce $PM_{2.5}$ pollution. We provide scientifically sound evidence that power plant emission controls are effective in reducing $PM_{2.5}$ pollution. Research that investigates the consequences of these regulations contributes critical evidence that EPA can use to defend current rules and consider future regulatory actions aimed at improving the public's health and the environment.

Finally, the research presented in this dissertation document investigates the impact of fracking industry on small scale $PM_{2.5}$ variability in PA. While we did not identify significant associations between fracking and $PM_{2.5}$ pollution in PA, we concluded that more air quality monitors are required to reach a substantiated conclusion about the particulate pollution risks associated with fracking. This research joins other recent publications in responding to the demand for an environmental assessment of fracking and contributes to the establishment of a basic understanding of the environmental health risks imposed by fracking activities (Davis 2012, Tollefson 2012, Carlton, Little et al. 2014, Rawlins 2014, Werner, Vink et al. 2015).

5.4 Innovations and Research Contributions

We expanded on the previously documented national trends of decreasing $PM_{2.5}$ pollution by investigating the changing airscape of $PM_{2.5}$ at smaller scales across the NE US. We applied statistically sound analytical techniques to investigate the impact of federal regulations on $PM_{2.5}$ pollution. We described a geostatistical-based approach for comparing before and after spatial surfaces that entails a spatially informed test of significance for small scale variations. In the absence of ample monitor data around the industrial activities associated with fracking, we utilized the methods we developed to predict $PM_{2.5}$ pollution at unmonitored locations across PA, thus completing a map of the PA airscape and allowing for the investigation of the influence of the fracking industry on $PM_{2.5}$ pollution. These results may be used to inform the placement of additional monitors, both in PA and

across the NE US. For example, the sparse monitoring of $PM_{2.5}$ in the northeastern region of PA, where fracking is well established, hindered our ability to detect associations between fracking and $PM_{2.5}$ trends

This research exemplifies analytical capabilities using limited resources. We used publicly available data and free, open-source software including the R statistical package and the geographic information system QGIS to complete the analyses in this dissertation. The accessibility of this research may foster its application to resource-limited agencies and researchers who may use similar techniques to investigate different aspects of air pollution.

The ability to complete a complex research investigation using publicly available data from different governmental sources may be a uniquely American talent. Dr. Hans Rosling of the Karolinska Institutet, remarking on the vast data collected and disseminated by the United States government, said, "It is US government at its best, without advocacy, providing facts that are useful for society and providing data free of charge on the internet for the world to use… When it comes to free data and transparency, United States of America is one of the best -- and that doesn't come easy from the mouth of a Swedish public health professor." (Rosling 2009) The accessibility of this free and open data creates an extraordinary avenue for public health engagement on the national level.

5.5 Future Directions

The methodology introduced in this dissertation document can be applied to new investigations of small scale spatial variations of pollutants across time. The

methods described in Chapter 3 may be applied in future research to determine whether new regulations attained a target decrease ($\mu g/m^3$) of ambient $PM_{2.5}$ concentrations.

Following the establishment of additional $PM_{2.5}$ monitors across PA, the research presented in Chapter 4 may be complemented by a large scale trend analysis, applying the methods described in Chapter 3 to PA. Additional monitors would allow for modeling of the large scale trends in $PM_{2.5}$ pollution in PA over a time period that includes increased regulations on $PM_{2.5}$ emissions and concurrent increases in fracking activity. The national policies identified in Chapter 3 that contribute to the reduction in $PM_{2.5}$ pollution across the NE US are, of course, actively impacting the $PM_{2.5}$ pollution in PA. Future reductions on $PM_{2.5}$ precursor emissions from power plants are expected from Phase II of the Cross-State Air Pollution Rule (CSAPR) in 2017 as well as from the Mercury and Air Toxics Standards (MATS), finalized by EPA on December 16, 2011 (United States Environmental Protection Agency 2016).

A comparison of the covariate coefficients from a model run before and after the establishment of fracking in a new area would allow us to investigate how the relationship between the covariates (the environmental determinants) and the outcome ($PM_{2.5}$) vary over a time period that includes the introduction of these increasing $PM_{2.5}$ regulations and the concurrent development of new fracking in PA. Further model analyses could include interaction terms to investigate how the relationship between model covariates and $PM_{2.5}$ pollution changes with the introduction of fracking. The large scale trend analysis would demonstrate whether

the establishment of fracking attenuates the positive impacts of federal regulations that reduce $PM_{2.5}$ pollution.

5.6 References

Berman, J. D., et al. (2015). "Evaluating methods for spatial mapping: Applications for estimating ozone concentrations across the contiguous United States." Environmental Technology & Innovation **3**: 1-10.

Bivand, R. S., et al. (2008). "Applied spatial data analysis with R. Springer."

Carlton, A. G., et al. (2014). "The data gap: can a lack of monitors obscure loss of clean air act benefits in fracking areas?" Environmental science & technology **48**(2): 893-894.

Cressie, N. (1993). "Statistics for spatial data: Wiley series in probability and statistics." Wiley-Interscience New York **15**: 16.

Davis, C. (2012). "The politics of "fracking": Regulating natural gas drilling practices in Colorado and Texas." Review of Policy Research **29**(2): 177-191.

Rawlins, R. A. (2014). "Planning for fracking on the Barnett shale: Urban air pollution, improving health based regulation, and the role of local governments." Retrieved 4/14/2016, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2410648.

Rosling, H. (2009). Let my dataset change your mindset. TED@State, TED.

Schabenberger, O. and C. A. Gotway (2005). Statistical methods for spatial data analysis, CRC press.

Tollefson, J. (2012). "Air sampling reveals high emissions from gas field." Nature **482**(7384): 139-140.

United States Environmental Protection Agency (2015, September 29). "About EPA: Our Mission and What We Do." Retrieved 4/13/2016, from https://www.epa.gov/aboutepa/our-mission-and-what-we-do.

United States Environmental Protection Agency (2016, February 23). "Mercury and Air Toxics Standards (MATS): Cleaner Power Plants." Retrieved 4/12/2016, from https://www3.epa.gov/mats/powerplants.html.

Werner, A. K., et al. (2015). "Environmental health impacts of unconventional natural gas development: A review of the current strength of evidence." Science of The Total Environment **505**: 1127-1141.

APPENDIX A

SUPPLEMENTAL MATERIALS FOR CHAPTER 2

Projections

The point-level variables were downloaded with their associated coordinate projections, and GIS was used to reproject these files as described below.

Data files of daily mean $PM_{2.5}$ monitor values downloaded from the EPA AQS website were associated with different coordinate systems. Of the N = 32,440 $PM_{2.5}$ monitor sites in our study, 17,637 monitors (54.37%) were associated with projection WGS84, 12,832 (39.56%) with projection NAD83, 1,511 (4.66%) with NAD27, and 460 monitors (1.42%) were listed as having an unknown projection coordinate system.

Data files of power plants locations did not specify a coordinate system as downloaded. Personal communication with EPA's Air Markets Program confirmed the power plant locations were associated with projection WGS84[1].

The geographic information system QGIS (version 2.10.1-Pisa) was used to reproject these data into the universal Transverse Mercator (UTM) coordinate system zone 18N (EPSG: 26918). UTM coordinates are expressed in meters, with higher X coordinate values associated with locations further east and higher Y coordinate values with locations further north. The X and Y coordinates were converted into km in the analyses.

Spatial trends

Spatial trends in monitor locations

[1] Personal communication with Kirk Nabors (nabors.kirk@epa.gov), 12/15/15

For the 32,440 monitors in the study, the average distance between the monitors is 399.60 km. The minimum distance between monitors is 0 km, which is expected due to sites with more than one monitor in the EPA AQS network, while the maximum distance is 1,652.42 km. For the 14 states in the study area, the average number of monitors per state from 2000 – 2014 is 2,317.14. Washington, D.C, contains the least monitors in our study (n = 531), and Pennsylvania contains the most (n = 5,361) (Table 1). There is a general trend of more monitors on the eastern side of the study area, closely following the I-95 corridor in the mid-Atlantic states (Figure 1). Heavier monitoring occurs in urban areas (Figure 1).

Spatial trends in $PM_{2.5}$ values

$PM_{2.5}$ daily averages show a general trend of higher values in the southern and western coordinates (Figure 3). The semivariogram of $PM_{2.5}$ displays spatial autocorrelation, indicating that $PM_{2.5}$ values are more similar for monitors located close together compared to monitors located further apart (Figure 4). Spatial trends were investigated by year and by season (Figures 5 & 6).

Tables and Figures

Descriptive data

State	Square miles in state	Number of counties in state	Number of monitors in state
Connecticut	5,543.41	8	2026
Delaware	2,488.72	3	1259
District Of Columbia	68.34	1	531
Maine	35,379.74	16	1884
Maryland	12,405.93	24	2836
Massachusetts	10,554.39	14	2857
New Hampshire	9,349.16	10	1225
New Jersey	8,722.58	21	3477
New York	54,554.98	62	4571
Pennsylvania	46,054.34	67	5361
Rhode Island	1,544.89	5	945
Vermont	9,616.36	14	809
Virginia	42,774.93	134	3093
West Virginia	24,230.04	55	1566

Table 1. Summary of states in NE US, including number of monitors in each state over the study period, 2000 – 2014.

Maps

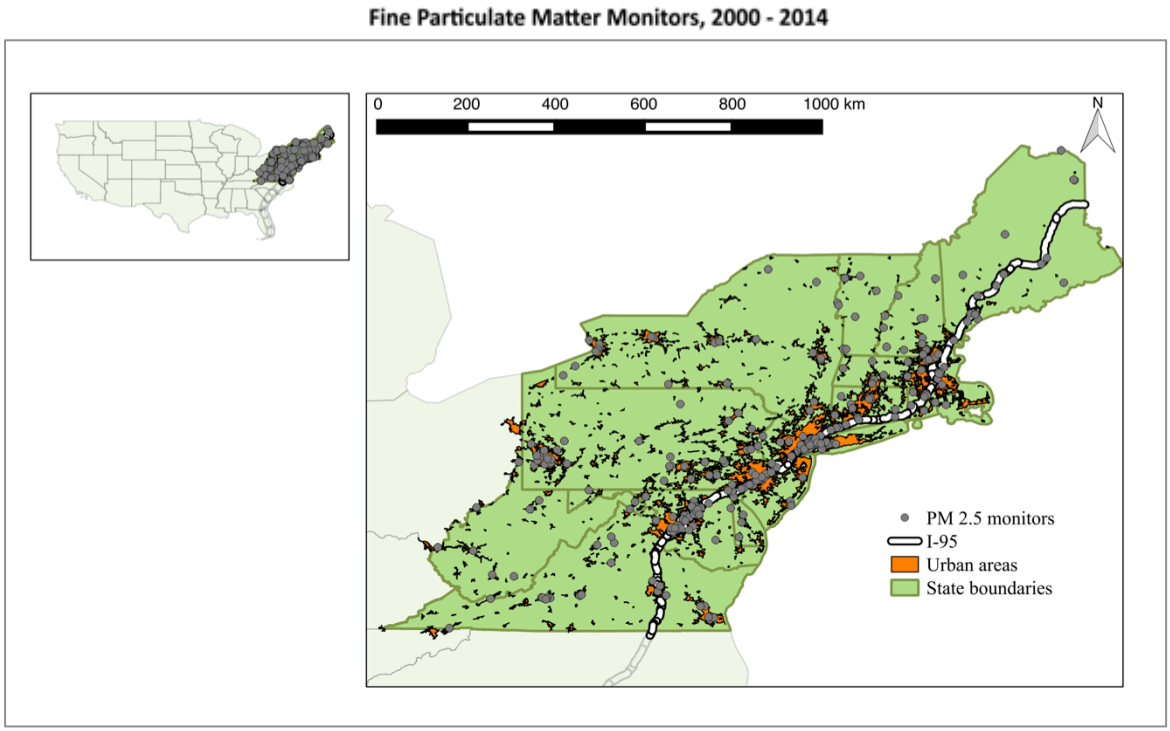

Figure 1. Map of $PM_{2.5}$ monitors in EPA AQS network, 2000 – 2014. "Urban areas" follow the 2010 Census urban areas definitions: "a densely settled core of census tracts and/or census blocks that meet minimum population density requirements, along with adjacent territory containing non-residential urban land uses as well as territory with low population density included to link outlying densely settled territory with the densely settled core. To qualify as an urban area, the territory identified according to criteria must encompass at least 2,500 people, at least 1,500 of which reside outside institutional group quarters." (United States Census Bureau 2015)

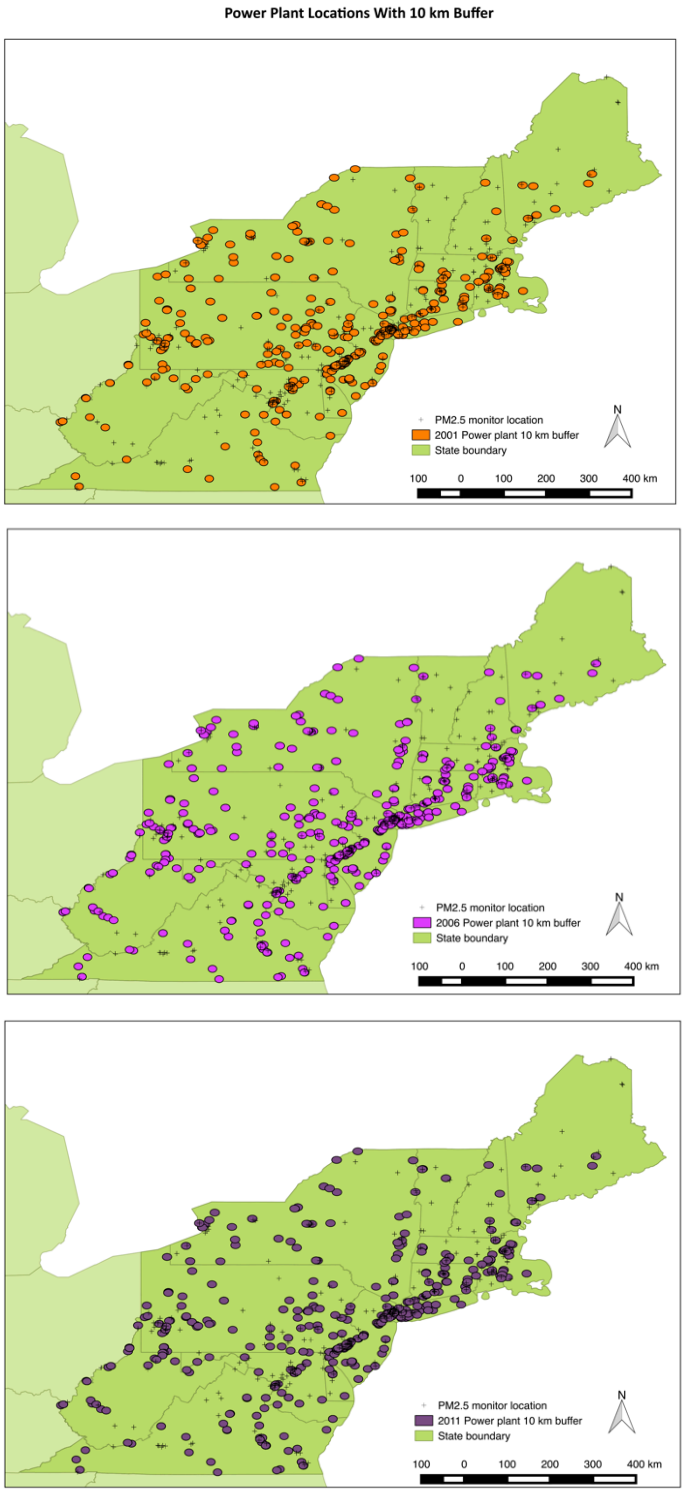

Figure 2. Power plant locations and buffers in 2001, 2006, and 2011.

Exploratory spatial analysis

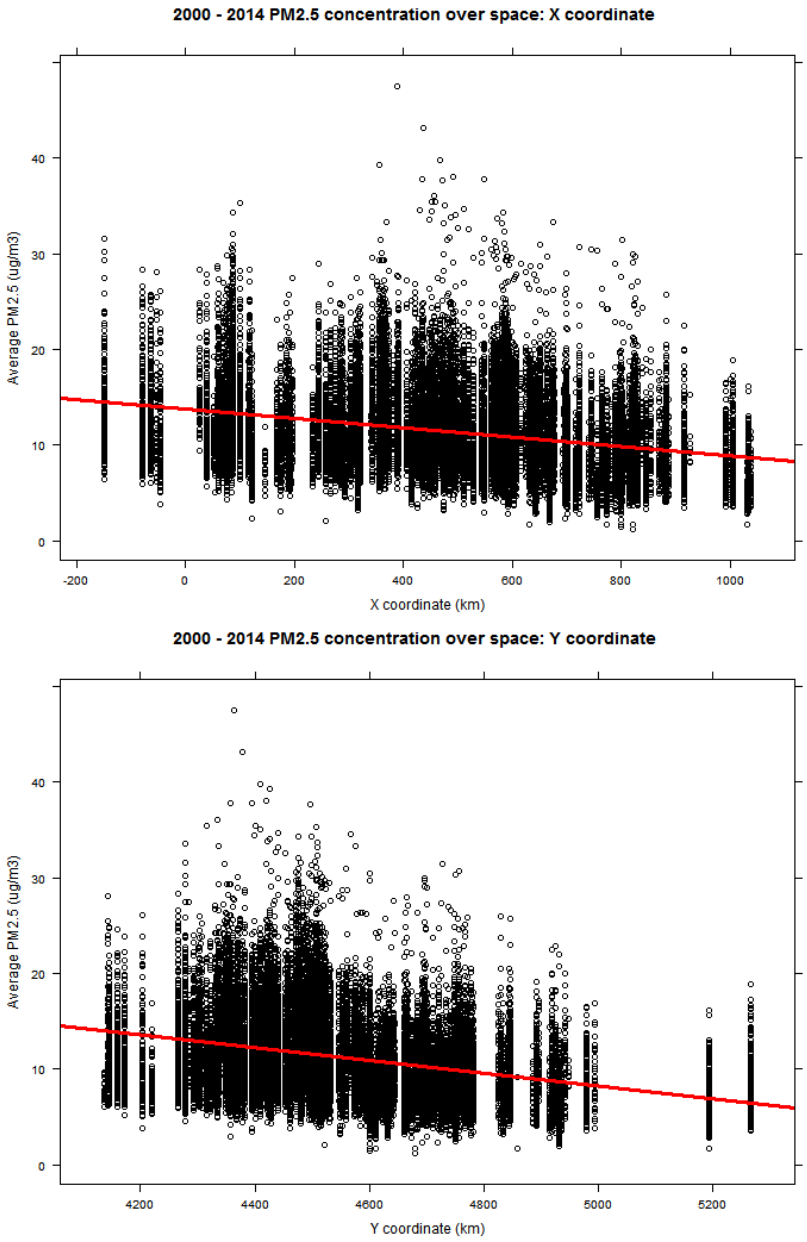

Figure 3. Scatterplot of average monthly $PM_{2.5}$ over coordinate values with trend lines. Coordinate values correspond to UTM zone 18N, moving from west to east with increasing X coordinate values and from south to north with increasing Y coordinate values.

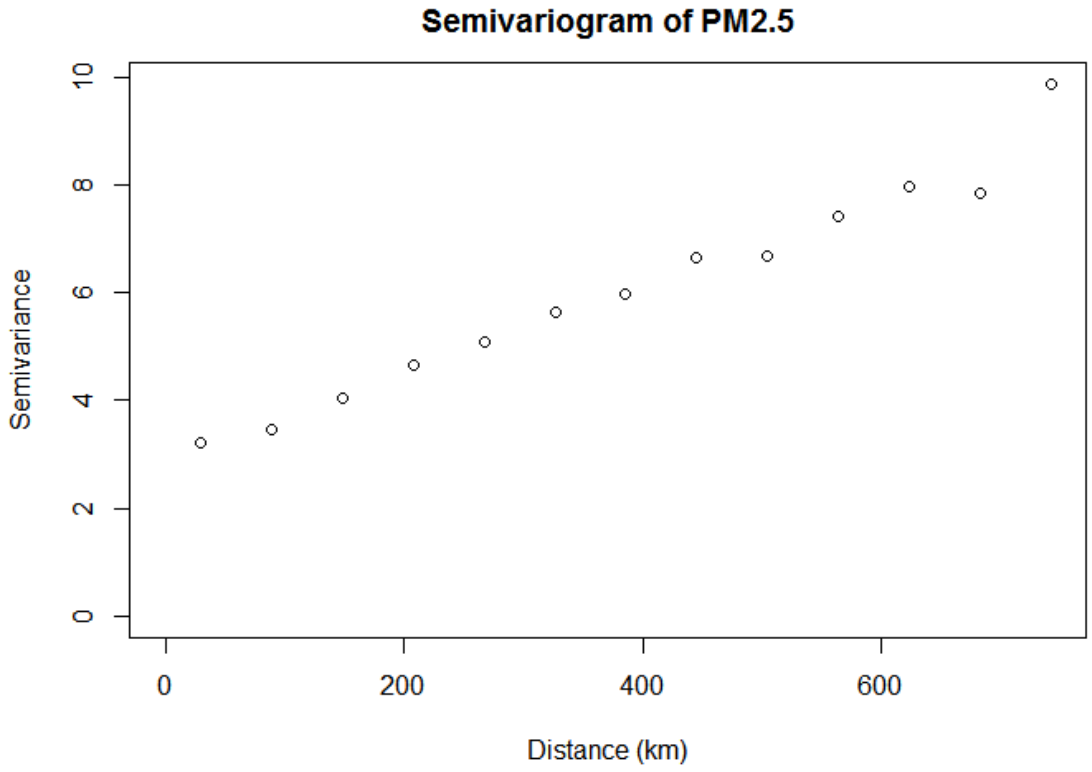

Figure 4. Semivariogram of the outcome, average $PM_{2.5}$, 2000 – 2014.

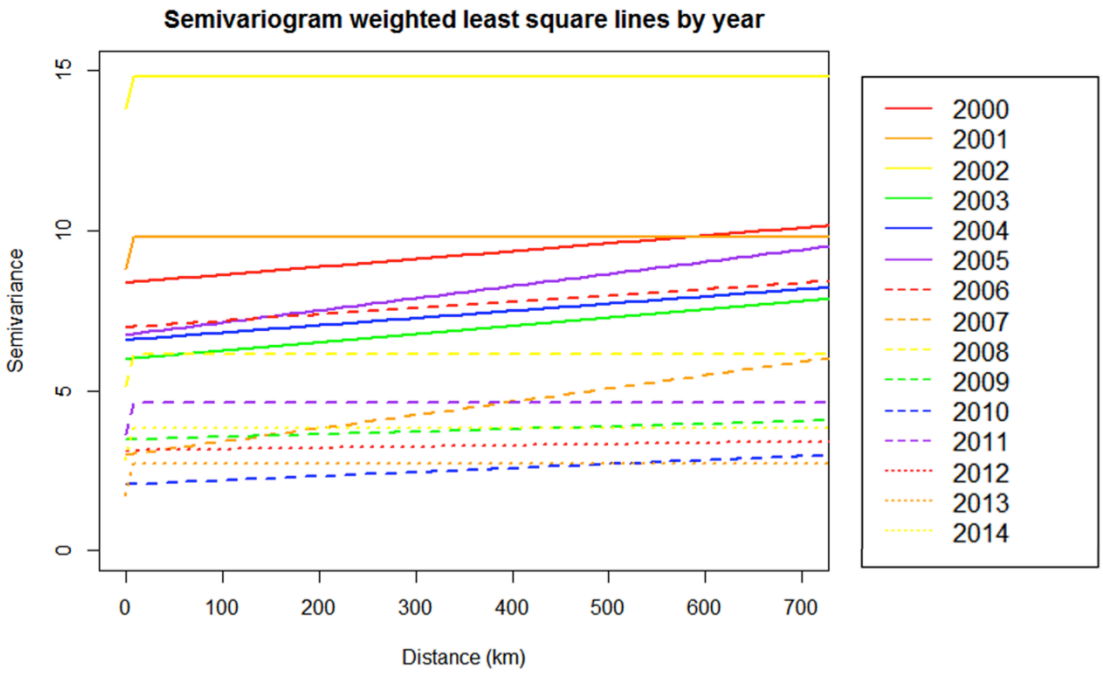

Figure 5. Weighted least squares line of the residual semivariograms of the final model by year. While the semivariance differs by year as evidenced by the Y axis intercepts, indicating that individual years have differing variability in the residuals, each year shows approximately a straight line in the weighted least squares line of the residual semivariogram.

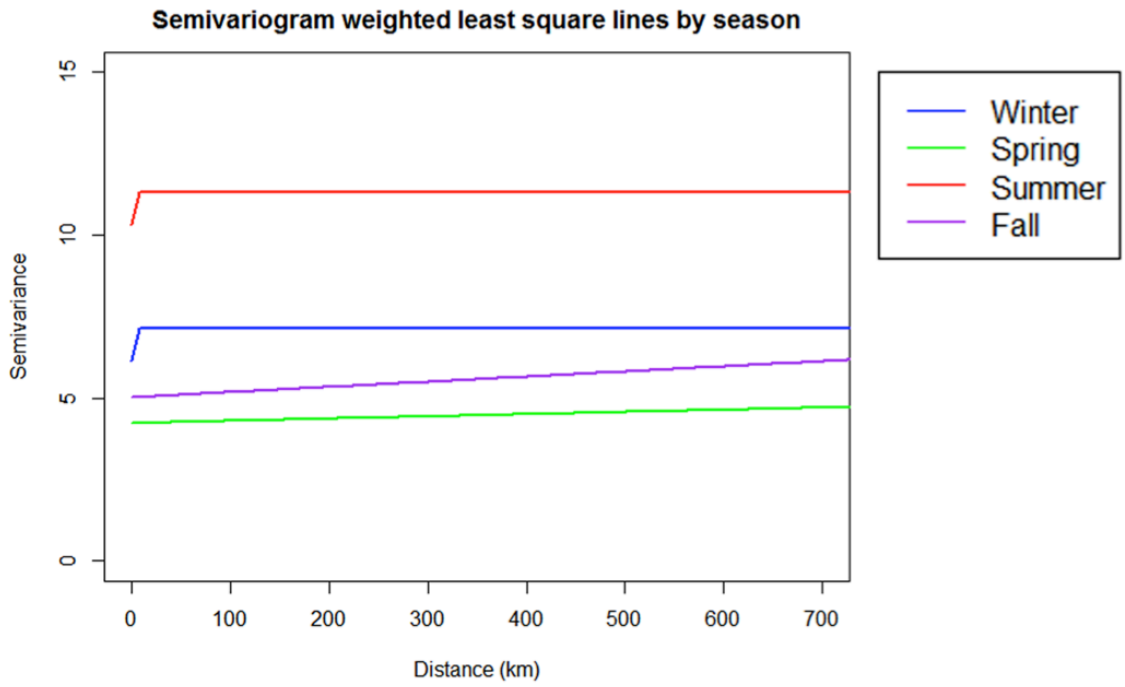

Figure 6. Weighted least squares line of the residual semivariograms of the final model by season. While the semivariance differs by season as evidenced by the Y axis intercepts, indicating that seasons have differing variability in the residuals, each season shows approximately a straight line in the weighted least squares line of the residual semivariogram.

Exploratory temporal analysis

Year	Number of monitors
2000	2246
2001	2330
2002	2385
2003	2167
2004	2148
2005	2169
2006	2021
2007	2086
2008	2119
2009	2109
2010	2145
2011	2148
2012	2108
2013	2114
2014	2145

Table 2. Summary of monitors per study year in the study area.

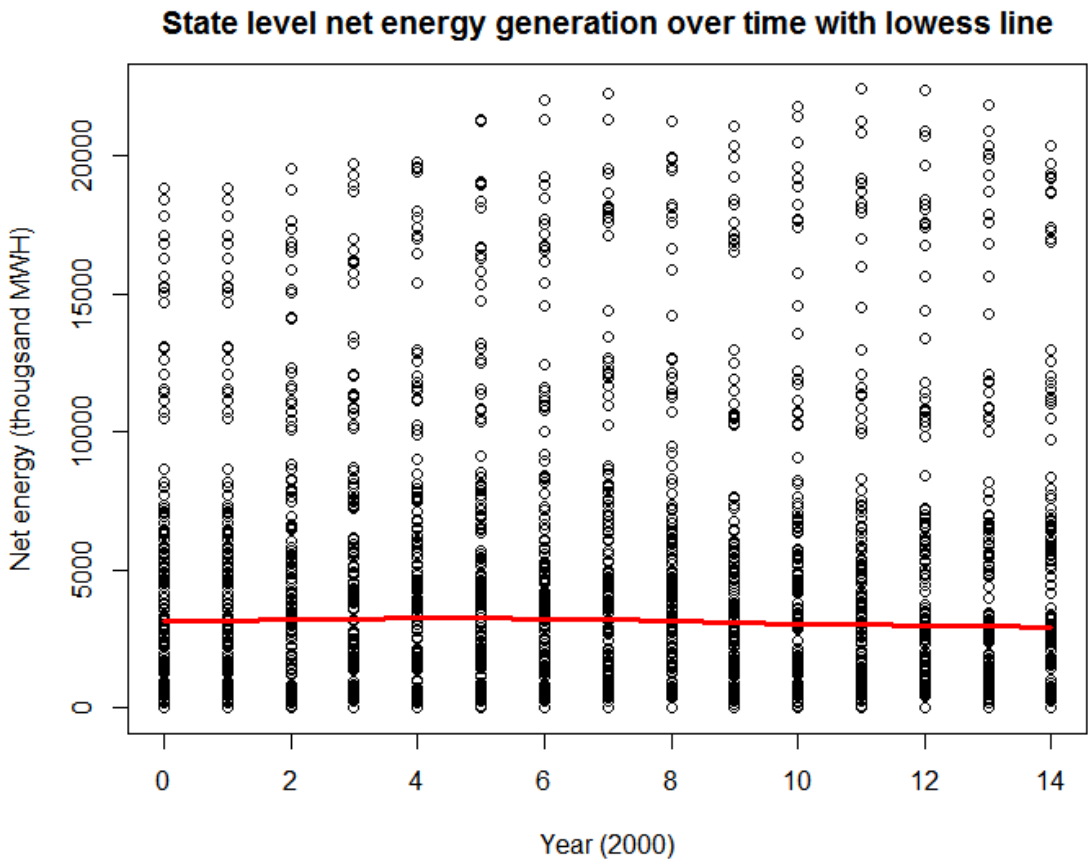

Figure 7. Scatterplot of state-level net energy generation, 2000 – 2014 (year 2000 is approximated using year 2001 data). The lowess line is fitted to the data so that trends may become apparent.

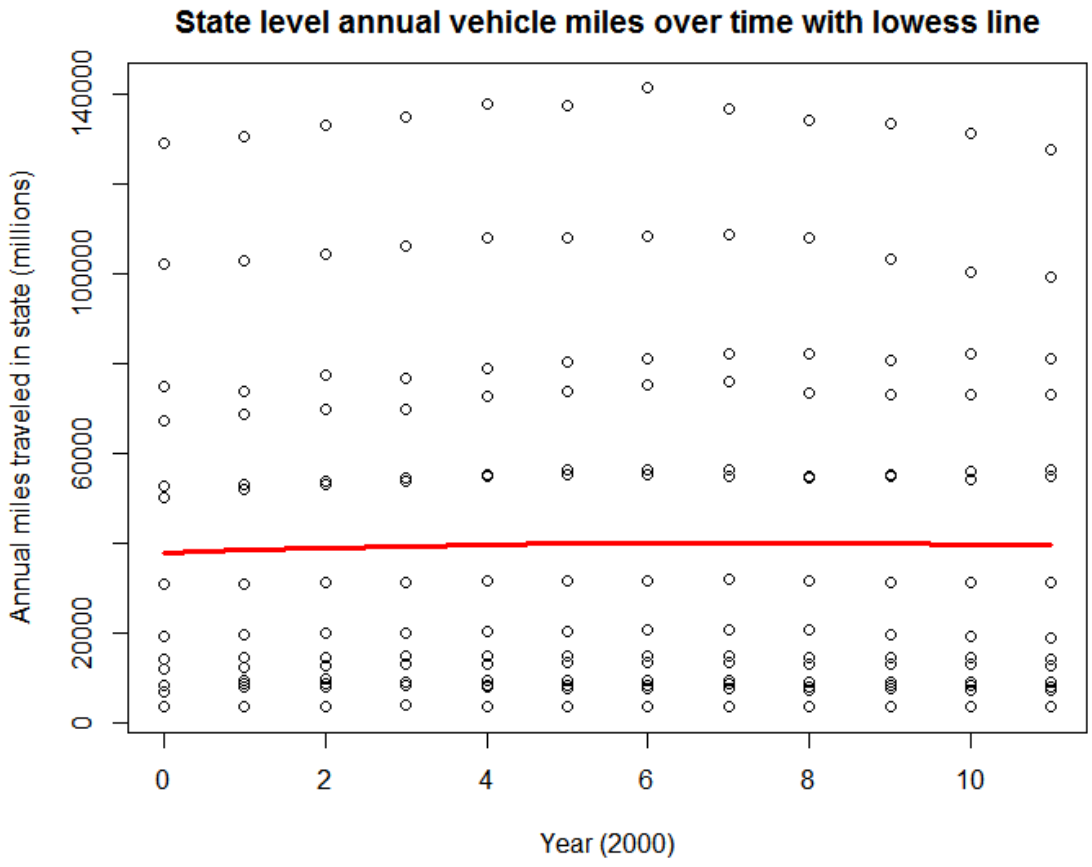

Figure 8. Scatterplot of state-level annual vehicle miles traveled, 2000 – 2011. The lowess line is fitted to the data to so that trends may become apparent.

Chi^2 distribution table

Degrees of Freedom	α =0.100	α = 0.050	α =0.025	α =0.010	α =0.005
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.21	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.86
5	9.236	11.07	12.833	15.086	16.75
6	10.645	12.592	14.449	16.812	18.548

Table 3. Chi^2 distribution table for select degrees of freedom (Duranczyk 2016). The table shows the critical values under different levels of significance (α).

References

Duranczyk, I. M. L., Suzanne; Stottlemyer, Janet (2016). "Chi-Square Distribution Table." Collaborative Statistics Using Spreadsheets. Retrieved 4/4/2016, from http://cnx.org/contents/ntrTKI5M@23.1:zkvpz8wr@2/Chi-Square-Distribution-Table.

United States Census Bureau (2015, February 9). "2010 Census Urban Area FAQs." Retrieved 3/13/2016, 2016, from https://www.census.gov/geo/reference/ua/uafaq.html.

APPENDIX B

SUPPLEMENTAL MATERIALS FOR CHAPTER 3

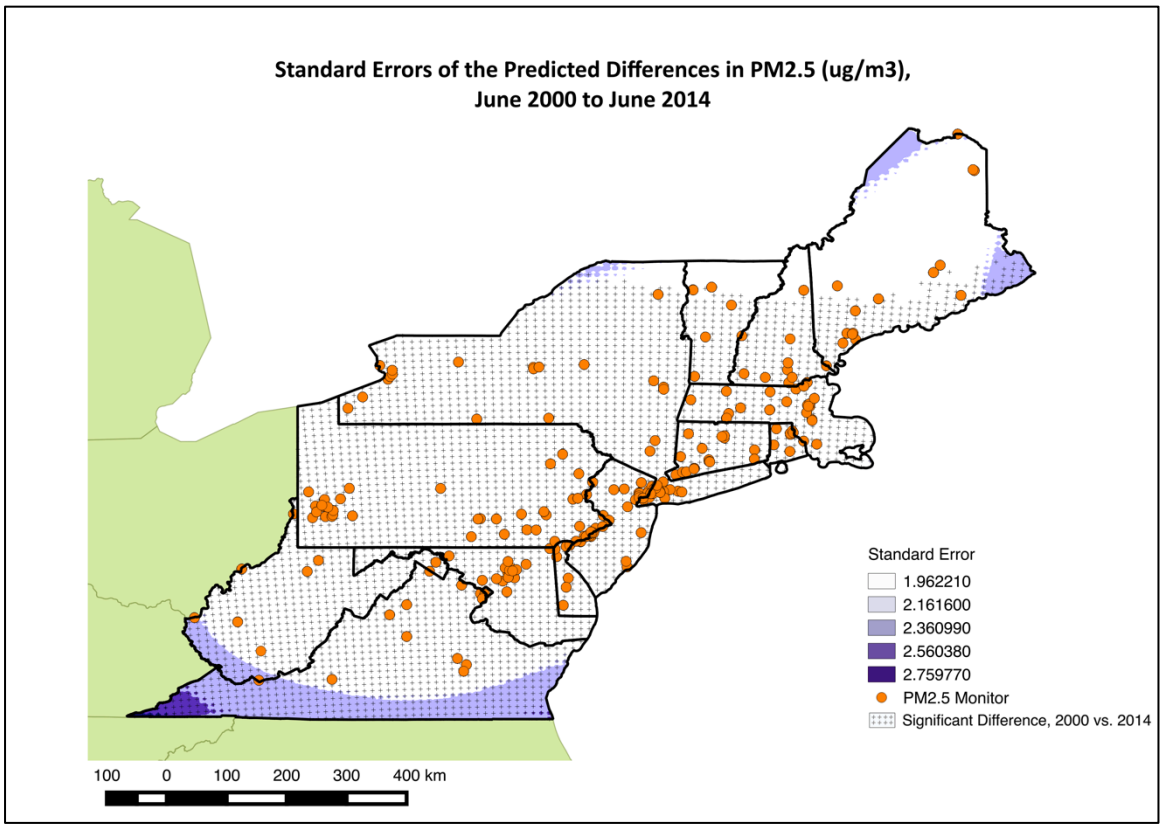

Figure 1. Map of the standard errors of the predicted differences in $PM_{2.5}$, June 2000 to June 2014. Significance is determined at $\alpha = 0.05$.

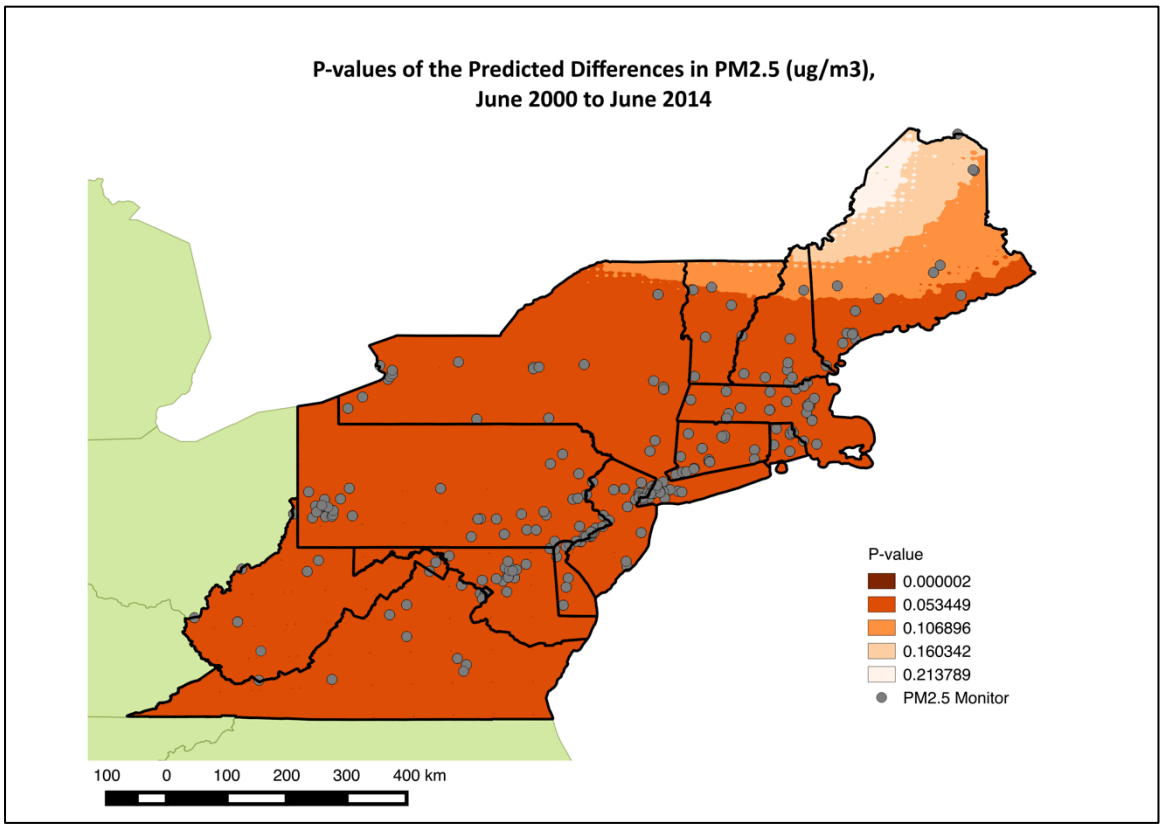

Figure 2. Map of the p-values indicating significance of the predicted differences in $PM_{2.5}$, June 2000 to June 2014.

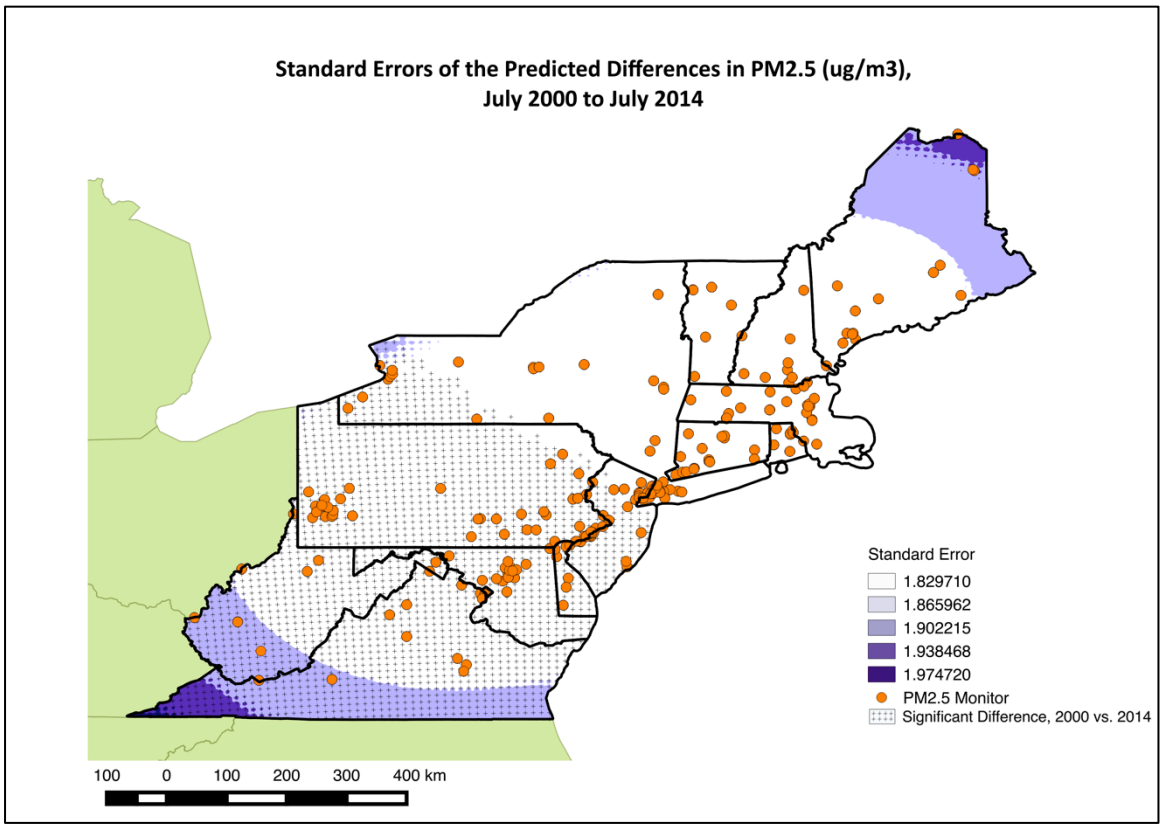

Figure 3. Map of the standard errors of the predicted differences in $PM_{2.5}$, July 2000 to July 2014. Significance is determined at $\alpha = 0.05$.

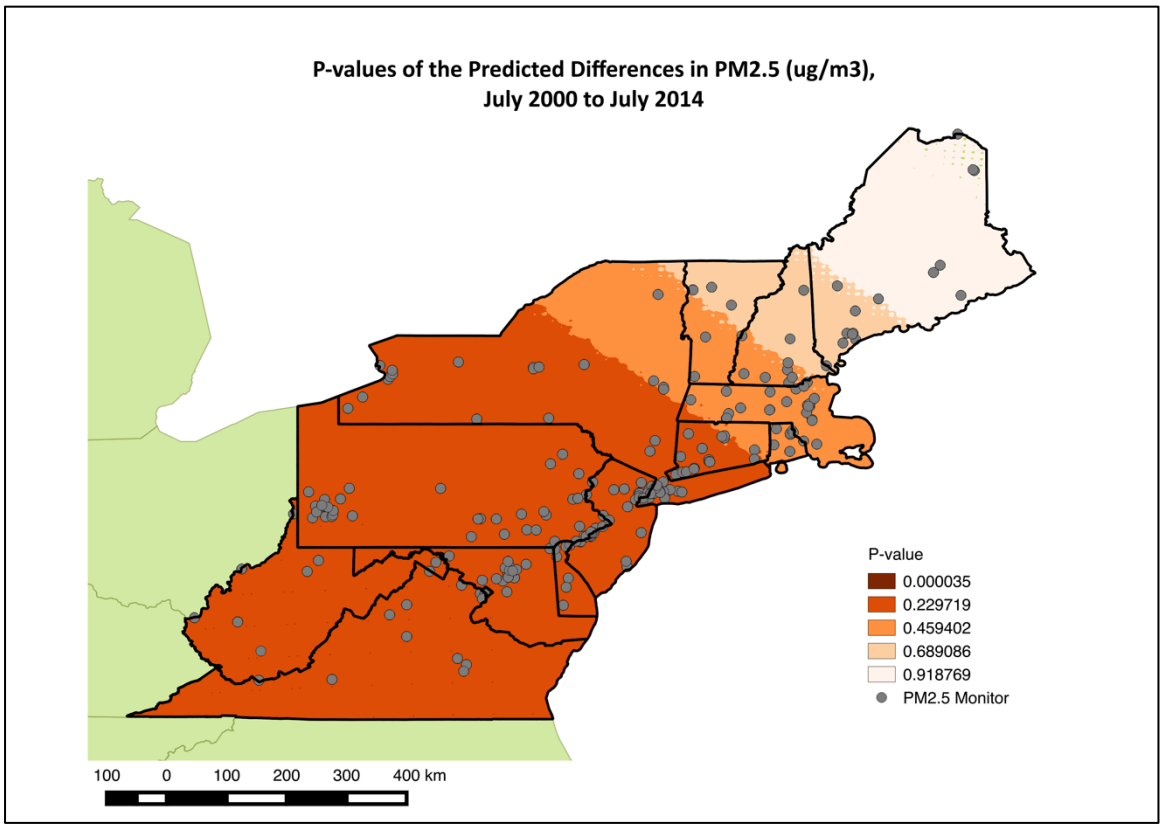

Figure 4. Map of the p-values indicating significance of the predicted differences in $PM_{2.5}$, July 2000 to July 2014.

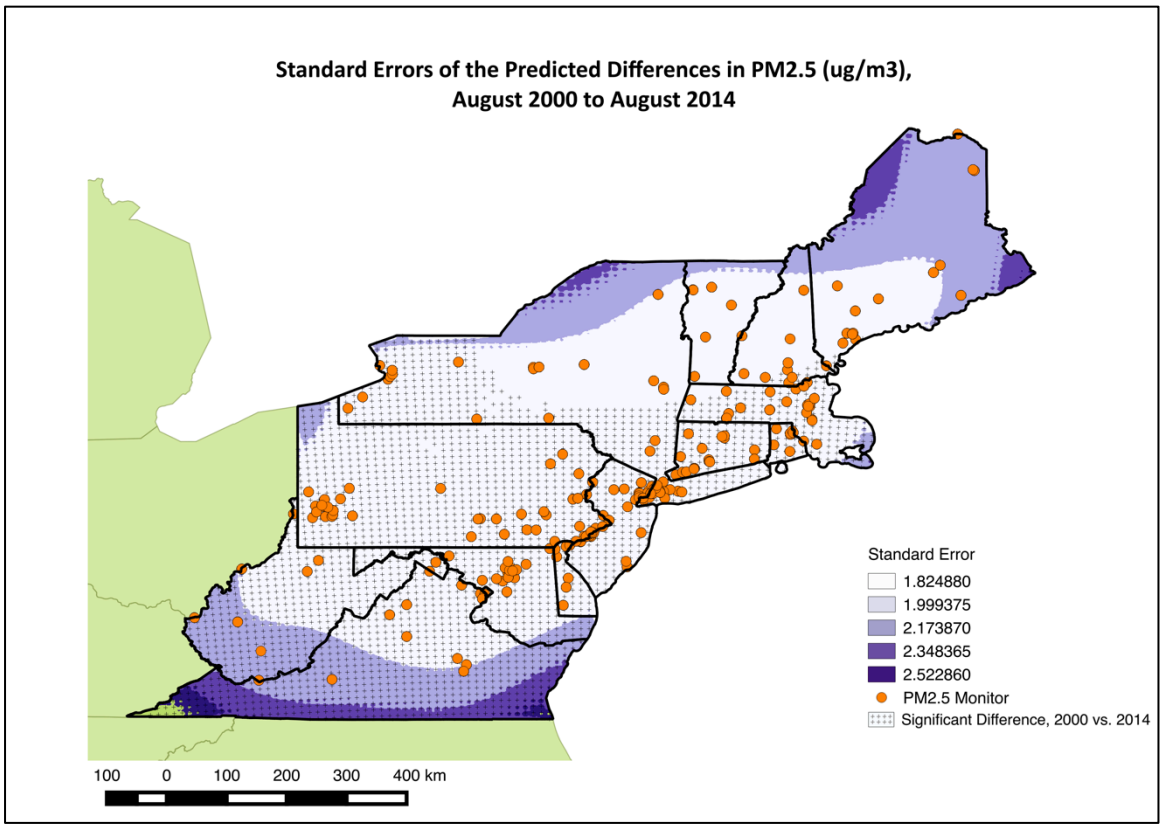

Figure 5. Map of the standard errors of the predicted differences in $PM_{2.5}$, August 2000 to August 2014. Significance is determined at $\alpha = 0.05$.

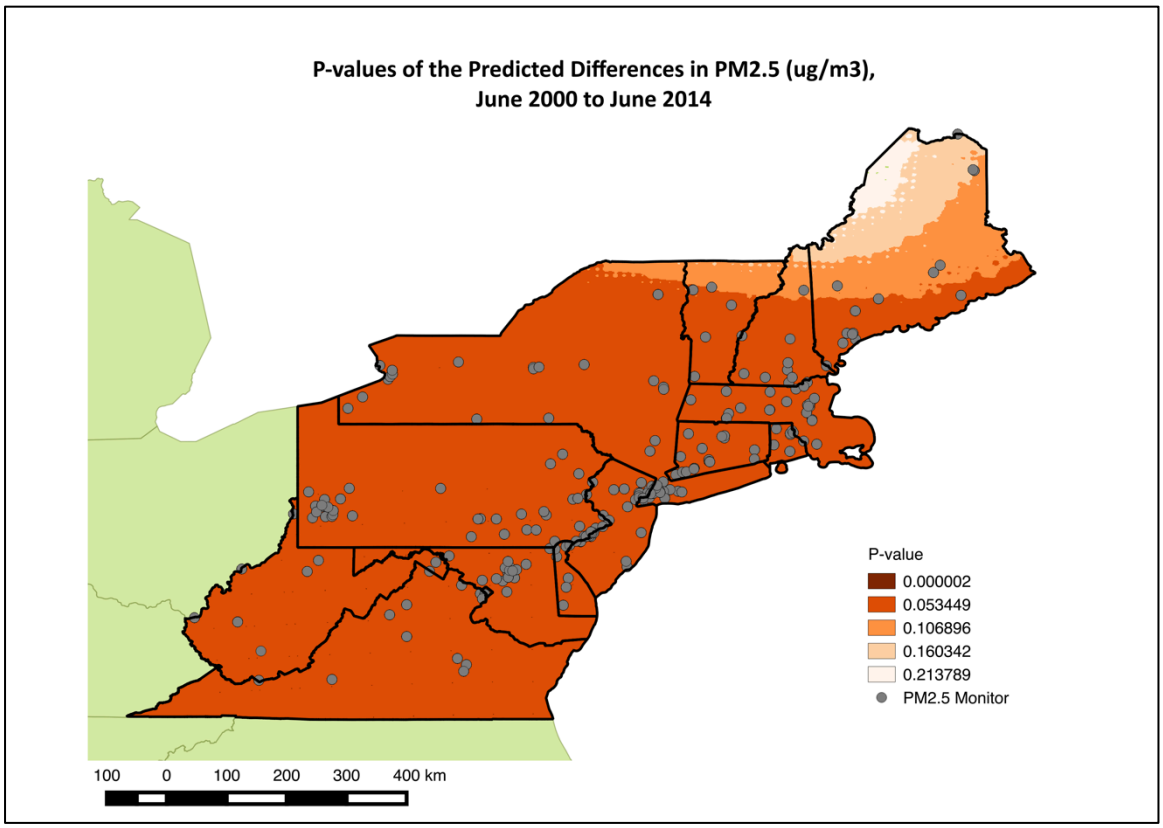

Figure 6. Map of the p-values indicating significance of the predicted differences in $PM_{2.5}$, August 2000 to August 2014.

APPENDIX C

SUPPLEMENTAL MATERIALS FOR CHAPTER 4

PA DEP Data

Data was downloaded from the Pennsylvania Department of Environmental Protection (DEP) oil and gas reporting website on 4/23/2016. Data definitions were obtained from the online data dictionary. We used data from DEP to ascertain two aspects of the fracking industry: where in PA is the majority of fracking occurring currently (2014), and when did the construction of new wells (spudding) reach a peak since fracking began in PA? Two datasets from DEP were considered in the analyses: the Wells Drilled by County Report provided the new wells data by year and the Oil and Gas Production Report provided data about active wells by year.

The Wells Drilled by County data was downloaded, and numbers for unconventional wells (fracking) were extracted. Using this dataset, 2011 showed the most fracking wells constructed in a single year during our study, with 1,619 wells (Chapter 4, Table 1).

Adding all of the new well construction data together from 2005 (the first year of fracking) until 2014 resulted in records for 7,571 new fracking wells built during this time (Chapter 4, Table 1). Accordingly, we expected to find less than or equal to 7,571 unconventional wells when we downloaded the 2014 active unconventional well data from DEP, since this data would include all of the new wells built from 2005 to 2014 minus those that are no longer active. The Gas Production Report requires users to download data in two separate files for 2014: January through June 2014, which reported 7,709 active unconventional wells, and July through December 2014, which reported 8,425 active unconventional wells. After duplicates were removed from the Gas Production Report (defined by

identical permit numbers), the new well data and the active well data numbers still did not align perfectly.

Our analysis of the fracking industry is an association based on location in which the relative abundance and general areas of fracking activity is more of a concern than obtaining the exact number of active wells. Accordingly, we used the latitude and longitude for the active unconventional well data from the Gas Production Report to indicate areas of fracking activity in Chapter 4, Figures 2 – 7, despite these small numerical discrepancies. These locations align with the counties identified as areas of new well construction and with the location of the Marcellus shale, supporting the use of this data in our analysis.

Figures

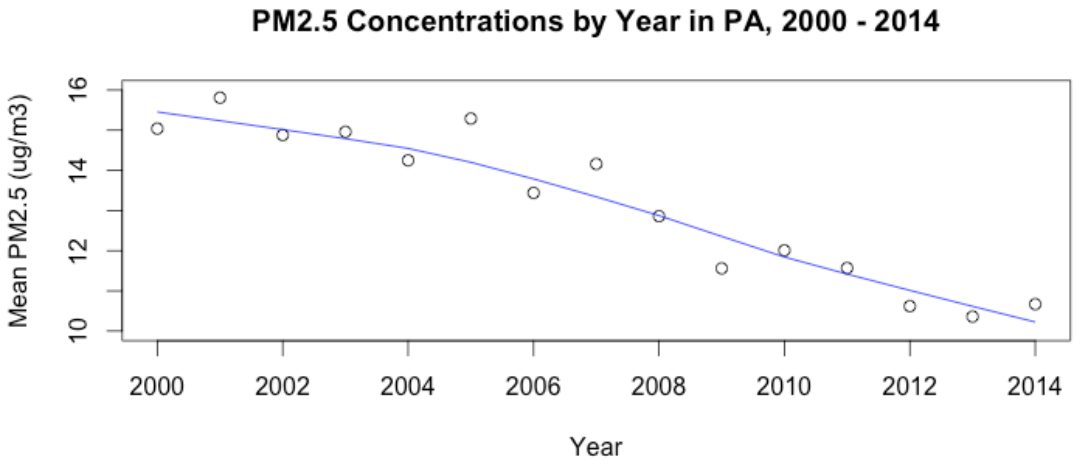

Figure 1. Mean $PM_{2.5}$ concentrations by year in PA, 2000 – 2014, with lowess line.

APPENDIX D: R-STATISTICAL SOFTWARE CODE

Introduction

This appendix contains the R statistical software code used in the analysis in chapters 2, 3, and 4 of this dissertation document. While this appendix does not report all of the code written and utilized, it provides examples of the pertinent functions and methods.

Exploratory Data Analysis (EDA)

```
require(plyr)
require(sp)
require(lattice)
require(latticeExtra)

names(PM)
nrow(PM)

summary(PM$Year)
summary(PM$AvgPM)

# How many PM monitors lie within pwr plant buffer
count(PM$X10km.buff)

# How many PM monitors per year / season
count <- count(PM$Year)
count <- count(PM$Season)

# Min and Max distance between monitors
dists <- dist(PM[,13:14])
max(dists)
min(dists)
mean(dists)
median(dist)

# States with highest and lowest monitors
count <- count(PM$State)
summary (count$freq)
count
```

```
# Any trends N, S, E, W?
coordinates(PM) <- c("X_meters", "Y_meters")

# Use "States" as object to investigate monitor placements
summary(PM$State)
spplot(PM, "State", do.log = T,
key.space=list(x=0.2,y=0.9,corner=c(0,1)),
scales=list(draw=T))

# How many monitor sites? Counties?
count(PM$SiteName)
count(PM$County)

# EDA: Spatial analysis of PM outcome
coordinates(PM) <- c("X_meters", "Y_meters")
spplot(PM, "AvgPM", do.log = T, colorkey = TRUE)
bubble(PM, "AvgPM", do.log = T, key.space = "bottom")

# Plot over Y in meters
xyplot(AvgPM ~ Y_meters, as.data.frame(PM), type = c("p"),
col=c("black"), lwd = 1, xlab = "Y coordinate (m)", ylab =
"Average PM2.5 (ug/m3)", main="1999 - 2014 PM concentration
over space: Y coordinate") +
  as.layer(xyplot(AvgPM ~ Y_meters, as.data.frame(PM), type
= c("r"), col=c("red"), lwd = 3))

# Plot over X in meters
xyplot(AvgPM ~ X_meters, as.data.frame(PM), type = c("p"),
col=c("black"), lwd = 1, xlab = "X coordinate (m)", ylab =
"Average PM2.5 (ug/m3)", main="1999 - 2014 PM concentration
over space: X coordinate") +
  as.layer(xyplot(AvgPM ~ X_meters, as.data.frame(PM), type
= c("r"), col=c("red"), lwd = 3))
```

Null and Bivariate Modeling

```
### MODEL: Null single level model (SLM)
nullSLM <- lm(AvgPM ~ 1, data = PM)

### MODEL: Bivariate model, PM and X and Y coordinates
m1<-lm(AvgPM~X_km)
m2<-lm(AvgPM~Y_km)
```

```
### MODEL: Bivariate model, PM and 10km power plant buffer
PM$X10km.buff <- factor(PM$X10km.buff)
m1<-lm(AvgPM~buff10km)

### MODEL: Bivariate model, PM and population
PMpop<-merge(PM,pop, by=c("State", "County", "Year"))

# Create pop per 1000
PMpop$PMPop1K <- PMpop$EstimatedPop / 1000

# Scale population by county size
PMpopco<-merge(PMpop,coarea, by=c("State", "County"))
PMpopco$Area.in.square.miles <-
as.numeric(PMpopco$Area.in.square.miles)
PMpopco$popco1k <- PMpopco$PMPop1K /
PMpopco$Area.in.square.miles

# Scatterplot to investigate a linear relationship between
the variables
pop<-PMpopco$popco1k
AvgPM<-PMpopco$AvgPM
plot(pop,AvgPM)

# linear regression model
m1<-lm(AvgPM~pop)

### MODEL: Bivariate model, PM and TRI
PMtri<-merge(PM,tri, by=c("State", "County", "Year"))

# create dummy variables
# Use quintile = 1 (the lowest TRI output level) as the
reference group
PMtri$tri2 <- PMtri$Quintile ==2
PMtri$tri3 <- PMtri$Quintile ==3
PMtri$tri4 <- PMtri$Quintile ==4
PMtri$tri5 <- PMtri$Quintile ==5

# linear regression model
m1<-lm(PMtri$AvgPM~PMtri$tri2 + PMtri$tri3 + PMtri$tri4 +
PMtri$tri5)

### MODEL: Bivariate model, PM and elevation
PMelv<-merge(PM,elv, by=c("State", "County"))
# linear regression model
m1<-lm(AvgPM~elv)

# MODEL: Bivariate model, PM AND ENERGY
```

```
PMEn<-merge(PM,EnMo, by=c("State", "Month", "Year"))

# Rescale energy, to hundred thousand megawatt hours
PMEn$Net_energy_thousandMWH100 <-
PMEn$Net_energy_thousandMWH / 100

# Scale energy by state area
PMEnst<-merge(PMEn,starea, by=("State"))
PMEnst$St.Area.in.square.miles <-
as.numeric(PMEnst$St.Area.in.square.miles)

PMEnst$enst100 <- PMEnst$Net_energy_thousandMWH100 /
PMEnst$St.Area.in.square.miles

energypermo<-PMEnst$enst100
AvgPM<-PMEnst$AvgPM

# Scatterplot to investigate a linear relationship between
the variables
plot(energypermo,AvgPM)

# linear regression model
m1<-lm(AvgPM~energypermo)

### MODEL: Bivariate model, PM and vehicle miles
PMVeh<-merge(PM,VehMi, by=c("State", "Year"))

# Create Billion Miles
PMVeh$BillMiles <- PMVeh$MillionMiles / 1000

# Scale vehicle miles by state area
PMVehst<-merge(PMVeh,starea, by=("State"))
PMVehst$St.Area.in.square.miles <-
as.numeric(PMVehst$St.Area.in.square.miles)

# Scatterplot to investigate a linear relationship between
the variables
VehicleMileYear<-PMVehst$BillMiles /
PMVehst$St.Area.in.square.miles
AvgPM<-PMVehst$AvgPM
plot(VehicleMileYear,AvgPM)

# linear regression model
m1<-lm(AvgPM~VehicleMileYear)

### MODEL: Bivariate model, PM and season
plot(season,AvgPM)
```

```
# linear regression model
m1<-lm(AvgPM~season)

### MODEL: Bivariate model, PM and year
plot(year, AvgPM)

# linear regression model
m1<-lm(AvgPM~year)
```

Multilevel Modeling

```
############################################################
Multilevel analysis - County nested within State Random
Effects
############################################################

require(pbkrtest)
require(car)
require(MASS)
require(lme4)
require(lmerTest)
require(nlme)

# Note: May run across following error when running lme
# from nlme package
# Error = nlminb problem, convergence error code = 1...
# iteration limit reached without convergence
# Fixed by following line:
ctrl <- lmeControl(opt='optim')


################## Build MLM #########################

#View normal distribution
AvgPM <- pm$AvgPM
qqnorm(AvgPM)
qqline(AvgPM)

# Since random effects are nested set REML to FALSE to use
# maximum likelihood

# Null single level model (SLM)
nullSLM <- lm(AvgPM ~ 1, data = pm)
summary(nullSLM)

# Null multilevel model (MLM)
# Run lme in package nlme to get p-values
```

```
nullMLM <- lme(AvgPM ~ 1, random = ~1 | State/County, data
= pm)
summary(nullMLM)

# Check out log likelihood of the models
# SLM
logLik(nullSLM)
AIC(nullSLM)
# MLM
logLik(nullMLM)
AIC(nullMLM)


###### Add in point level variables #####

MLMpt <- lme (AvgPM ~ X10km.buff + X_km + Y_km, random = ~1
| State/County, data = pm)
summary(MLMpt)

AIC(MLMpt)
logLik(MLMpt)


###### Add in county variables #####

MLMco <- lme (AvgPM ~ X10km.buff + X_km + Y_km +
Mean.Elevation..m. + tri2 + tri3 + tri4 + tri5 + popco1K,
random = ~1 | State/County, data = pm)
summary(MLMco)

AIC(MLMco)
logLik(MLMco)


##### Add in state level variables #####

MLMst <- lme (AvgPM ~ X10km.buff + X_km + Y_km +
Mean.Elevation..m. + tri2 + tri3 + tri4 + tri5 + popco1K +
vehstbillmi + enst100, random = ~1 | State/County, data =
pm)
summary(MLMst)

AIC(MLMst)
logLik(MLMst)


#### Add in temporal variables #####
```

```
MLMall <- lme (AvgPM ~ X10km.buff + X_km + Y_km +
Mean.Elevation..m. + tri2 + tri3 + tri4 + tri5 + popco1K +
vehstbillmi + enst100 + Season + Year, random = ~1 |
State/County, data = pm)
summary(MLMall)

AIC(MLMall)
logLik(MLMall)


###### Compare ensemble model without random effect to the
# MLMall #####

# ensemble SLM
SLMall <- lm (AvgPM ~ X10km.buff + X_km + Y_km +
Mean.Elevation..m. + tri2 + tri3 + tri4 + tri5 + popco1K +
vehstbillmi + enst100 + Season + Year, data = pm)
summary(SLMall)

AIC(SLMall)
logLik(SLMall)


###### Ensemble to final model ######
# Recall best model is MLM with random effect on County
# nested within State

MLMall <- lme (AvgPM ~ X10km.buff + X_km + Y_km +
Mean.Elevation..m. + tri2 + tri3 + tri4 + tri5 + popco1K +
vehstbillmi + enst100 + Season + Year, random = ~1 |
State/County, data = pm)
summary(MLMall)


# Covariates that are not significant (p >= 0.05): X_km,
# popco1K, TRI
# Spring is also not significant but will keep in final
# model bc is germane to the season covariate

##### Final model, only significant covariates #####
# Nested random effects model
# Allows for random intercepts for each county within
# states

MLMfinal <- lme (AvgPM ~ X10km.buff + Y_km +
Mean.Elevation..m. + vehstbillmi + enst100 + Season + Year,
random = ~1 | State/County, data = pm)
summary(MLMfinal)
```

Spatial Analysis of Multilevel Model

```
require (nlme)
ctrl <- lmeControl(opt='optim')
require(lme4)
require(lmerTest)
require (lattice)
require(latticeExtra)
require (geoR)
require (gstat)
require (maptools)
gpclibPermit()
require (PBSmapping)
require (sp)

#############################################################
 Spatial analysis of step-wise MLM
#############################################################

# Here we explore whether adding covariates (building from
# the null SLM to ensemble MLM) accounts for the
# residual spatial variation
# We will investigate the covariates from the final MLM:
# MLMfinal <- lme (AvgPM ~ X10km.buff + Y_km +
# Mean.Elevation..m. + vehstbillmi + enst100 + Season +
# Year, random = ~1 | State/County, data = pm)

# Null model SLM
nullSLM <- lm(AvgPM ~ 1, data = pm)
nullSLMres.geodata <-as.geodata(cbind(pm$X_km, pm$Y_km,
nullSLM$residuals), rep.data.action = mean)
# "rep.data.action = mean" to use mean of all values at
# same point (mean of all years PM at single monitor)

# Plot the variogram of the residuals
max(dist(nullSLMres.geodata$coords))
# 1542.91 km
v0<-variog(nullSLMres.geodata,max.dist=1542.91/2)
plot(v0, xlab="Distance (km)", ylab="Semivariance")
title("Semivariogram, null model",line=1)
# Export figure w/o wls line
# Eyeball: Gaussian, range = 600, sill = 8, nugget = 3
# Weighted least squares fit of the variogram v0
# Gaussian
```

```
v0.wlsfit1<-
variofit(v0,ini.cov.pars=c(8,600),nugget=3,cov.model="gauss
ian")
lines(v0.wlsfit1)
# Export figure

###### Add in point level variables #####

# No random effect
ptSLM <- lm (AvgPM ~ X10km.buff + Y_km, data = pm)
ptSLMres.geodata <-as.geodata(cbind(pm$X_km, pm$Y_km,
ptSLM$residuals), rep.data.action = mean)

# Plot the variogram of the residuals
max(dist(ptSLMres.geodata$coords))
# 1542.91 km
v1<-variog(ptSLMres.geodata,max.dist=1542.91/2)
plot(v1, xlab="Distance (km)", ylab="Semivariance")
title("Semivariogram, point-level covariate model",line=1)
# Export figure
# Eyeball: Gaussian, range = 600, sill = 4, nugget = 3
# Weighted least squares fit of the variogram v1
# Gaussian
v1.wlsfit1<-
variofit(v1,ini.cov.pars=c(4,600),nugget=3,cov.model="gauss
ian")
lines(v1.wlsfit1)
v1.wlsfit2<-
variofit(v1,ini.cov.pars=c(4,600),nugget=3,cov.model="expon
ential")
lines(v1.wlsfit2)
# export figure

###### Add in county variables

# No random effect
coSLM <- lm (AvgPM ~ X10km.buff + Y_km + Mean.Elevation..m.,
data = pm)
coSLMres.geodata <-as.geodata(cbind(pm$X_km, pm$Y_km,
coSLM$residuals), rep.data.action = mean)

# Plot the variogram of the residuals
max(dist(coSLMres.geodata$coords))
# 1542.91 km
v2<-variog(coSLMres.geodata,max.dist=1542.91/2)
plot(v2, xlab="Distance (km)", ylab="Semivariance")
title("Semivariogram, county-level covariate model",line=1)
# Export figure
```

```
# Eyeball: Gaussian, range = 600, sill = 4, nugget = 3
# Weighted least squares fit of the variogram v1
# Gaussian
v2.wlsfit1<-
variofit(v2,ini.cov.pars=c(4,600),nugget=3,cov.model="gauss
ian")
lines(v2.wlsfit1)
# Exp
v2.wlsfit2<-
variofit(v2,ini.cov.pars=c(4,600),nugget=3,cov.model="expon
ential")
lines(v2.wlsfit2)
# Export figure

# Add county-level random effect
coMLM <- lme(AvgPM ~ X10km.buff + Y_km + Mean.Elevation..m.,
random = ~1 | County, data = pm)
coMLMres.geodata <-as.geodata(cbind(pm$X_km, pm$Y_km,
coMLM$residuals), rep.data.action = mean)

# Plot the variogram of the residuals
max(dist(coMLMres.geodata$coords))
# 1542.91 km
v3<-variog(coMLMres.geodata,max.dist=1542.91/2)
plot(v3, xlab="Distance (km)", ylab="Semivariance")
title("Semivariogram, county-level random effect
model",line=1)
# Export figure
# Eyeball: Gaussian, range = 600, sill = 4, nugget = 3
# Weighted least squares fit of the variogram v1
# Gaussian
v3.wlsfit1<-
variofit(v3,ini.cov.pars=c(4,600),nugget=3,cov.model="gauss
ian")
lines(v3.wlsfit1)
# Exp
v3.wlsfit2<-
variofit(v3,ini.cov.pars=c(4,600),nugget=3,cov.model="expon
ential")
lines(v3.wlsfit2)
# Export figure

##### Add in state level variables #####

# Without state level random effect
stMLMnoRE <- lme (AvgPM ~ X10km.buff + Y_km +
Mean.Elevation..m. + vehstbillmi + enst100, random = ~1 |
County, data = pm)
```

```
stMLMnoREres.geodata <-as.geodata(cbind(pm$X_km, pm$Y_km,
stMLMnoRE$residuals), rep.data.action = mean)

# Plot the variogram of the residuals
max(dist(stMLMnoREres.geodata$coords))
# 1542.91 km
v4<-variog(stMLMnoREres.geodata,max.dist=1542.91/2)
plot(v4, xlab="Distance (km)", ylab="Semivariance")
title("Semivariogram, state-level covariate model",line=1)
# Export figure
# Weighted least squares fit of the variogram
# Gaussian
v4.wlsfit1<-
variofit(v4,ini.cov.pars=c(4,600),nugget=3,cov.model="gauss
ian")
lines(v4.wlsfit1)
v4.wlsfit2<-
variofit(v4,ini.cov.pars=c(4,600),nugget=3,cov.model="expon
ential")
lines(v4.wlsfit2, col = "red")
v4.wlsfit3<-
variofit(v4,ini.cov.pars=c(4,600),nugget=3,cov.model="spher
ical")
lines(v4.wlsfit3, col = "blue")
# Note that lines don't change with changing model types
(gaussian, exponential, spherical)
# Export figure with black line only

# With state level random effect
# adding in season allowed convergence
stMLMre <- lme (AvgPM ~ X10km.buff + Y_km +
Mean.Elevation..m. + vehstbillmi + enst100 + Season, random
= ~1 | State/County, data = pm)
stMLMreres.geodata <-as.geodata(cbind(pm$X_km, pm$Y_km,
stMLMre$residuals), rep.data.action = mean)
# Plot the variogram of the residuals
max(dist(stMLMreres.geodata$coords))
# 1542.91 km
v5<-variog(stMLMreres.geodata,max.dist=1542.91/2)
plot(v5, xlab="Distance (km)", ylab="Semivariance")
title("Semivariogram, state-level random effect
model",line=1)
# Export figure
# Eyeball: Gaussian, range = 400, sill = 5, nugget = 5
# Weighted least squares fit of the variogram
# Gaussian
```

```
v5.wlsfit1<-
variofit(v5,ini.cov.pars=c(5,400),nugget=5,cov.model="gauss
ian")
lines(v5.wlsfit1)
# Exp
v5.wlsfit2<-
variofit(v5,ini.cov.pars=c(5,400),nugget=5,cov.model="expon
ential")
lines(v5.wlsfit2)
# Export figure

#### Add in temporal variables to arrive at final MLM #####

MLMfinal <- lme (AvgPM ~ X10km.buff + Y_km +
Mean.Elevation..m. + vehstbillmi + enst100 + Season + Year,
random = ~1 | State/County, data = pm)
MLMfinalres.geodata <-as.geodata(cbind(pm$X_km, pm$Y_km,
MLMfinal$residuals), rep.data.action = mean)

# Plot the variogram of the residuals
max(dist(MLMfinalres.geodata$coords))
# 1542.91 km
v6<-variog(MLMfinalres.geodata,max.dist=1542.91/2)
plot(v6, xlab="Distance (km)", ylab="Semivariance")
title("Semivariogram, final multilevel model",line=1)
# Export figure
# Eyeball: Gaussian, range = 300, sill = 15, nugget = 10
# Weighted least squares fit of the variogram
# Gaussian
v6.wlsfit1<-
variofit(v6,ini.cov.pars=c(15,300),nugget=10,cov.model="gau
ssian")
lines(v6.wlsfit1)
# Exp
v6.wlsfit2<-
variofit(v6,ini.cov.pars=c(15,300),nugget=10,cov.model="exp
onential")
lines(v6.wlsfit2)
# Export figure


############################################################
Spatial analysis of final MLM by year
############################################################
# Here we explore whether the spatial analysis from the
# final MLM as explored above holds when we explore each
# year separately
```

```
###### 2000 #####

# Subset year pm data
pm00 <- subset(pm, Year==0)
str(pm00)
# Set buffer as factor
pm00$X10km.buff <- factor(pm00$X10km.buff)
str(pm00)

names(pm00)
pm00$Xkm <- pm00$X_meters/1000
pm00$Ykm <- pm00$Y_km

# Final model for subset (remove year)
m0 <- lme (AvgPM ~ X10km.buff + Y_km + Mean.Elevation..m. +
vehstbillmi + enst100 + Season, random = ~1 | State/County,
data = pm00)
summary(m0)

# Take the residuals of the model & make a geodata object
m0r <- resid(m0)
m0res.geodata <-as.geodata(cbind(pm00$Xkm, pm00$Ykm, m0r ))
max(dist(m0res.geodata$coords))
# 1542.82

# Variogram of residuals
v0<-variog(m0res.geodata,max.dist=1542.82/2)
plot(v0, xlab="Distance (km)", ylab="Semivariance")
title("Semivariogram, final MLM 2000",line=1)
# Export

# Weighted least squares fit of the variogram
v0.wlsfit1<-
variofit(v0,ini.cov.pars=c(10,600),nugget=7,cov.model="expo
nential")
lines(v0.wlsfit1)

#### Above code repeated for every year in the dataset ####


############################################################
Spatial analysis of final MLM by season
############################################################

#### Fall ####
```

```
# Subset season pm data
pm_fall <- subset(pm, Season=="Fall")
str(pm_fall)
# Set buffer as factor
pm_fall$X10km.buff <- factor(pm_fall$X10km.buff)
str(pm_fall)

pm_fall$Xkm <- pm_fall$X_meters/1000
pm_fall$Ykm <- pm_fall$Y_km

# Final model for subset (remove season)
m_fall <- lme (AvgPM ~ X10km.buff + Y_km +
Mean.Elevation..m. + vehstbillmi + enst100 + Year, random =
~1 | State/County, data = pm_fall)
summary(m_fall)

# Take the residuals of the model & make a geodata object
m_fallr <- resid(m_fall)
m_fallres.geodata <-as.geodata(cbind(pm_fall$Xkm,
pm_fall$Ykm, m_fallr ))
max(dist(m_fallres.geodata$coords))
# 1542.82

# Variogram of residuals
v_fall<-variog(m_fallres.geodata,max.dist=1542.82/2)
plot(v_fall, xlab="Distance (km)", ylab="Semivariance")
title("Semivariogram, final MLM, fall",line=1)

# Weighted least squares fit of the variogram
v_fall.wlsfit1<-
variofit(v_fall,ini.cov.pars=c(4,600),nugget=3.5,cov.model=
"exponential")
lines(v_fall.wlsfit1)

#### Above code repeated for Spring, Summer, Winter ####


############ Conclusion ###################################

# The final MLM accounts for residual spatial variation
# Further analysis will be done using the MLM with errors ~
# N(o, sigma squared)
```

Large Scale Trend Analysis

```
# Aim 2, large scale trend analysis
```

```
################## Stratified analysis, 2000 and 2014 ###

############ 2000 ####

#MLM packages
require(lme4)
require(lmerTest)
require(nlme)

# Read in data file of pm + covariates
pm<-read.csv("PMelvTriPopcoEnstVehst.csv")

# Subset year pm
pm00 <- subset(pm, Year==0)
str(pm00)
# Set buffer as factor
pm00$X10km.buff <- factor(pm00$X10km.buff)
str(pm00)

# Average PM in 2000
summary(pm00$AvgPM)

# Final model for pm year (remove year covariate)
MLMfinal00 <- lme (AvgPM ~ X10km.buff + Y_km +
Mean.Elevation..m. + vehstbillmi + enst100 + Season, random
= ~1 | State/County, data = pm00)
summary(MLMfinal00)

### Model validation: Predictive power of model via out-of-
# sample cross validation ###

# Predict 10% without replacement
nrow(pm00)
# PM 00 = 2044
# 10% of 2044 = 204.4
# Predict at 204 sample points
sub00 <- pm00[sample(nrow(pm00), 204),]
nrow(sub00)
head(sub00, 10)

# Predict at the 204 sample points
# Create column for predicted results
sub00$predictPM <- predict(MLMfinal00, sub00, level = 0:1)

# View results
names(sub00)
head(sub00, 10)
tail(sub00, 10)
```

```
# Note you have to "unnest" county and state in order to
move forward in this code
# Use LinearizeNestedList function
library(devtools)
source_gist(4205477) #loads the function

sub00f <- LinearizeNestedList(sub00, LinearizeDataFrames =
TRUE)
names(sub00f)

sub00f <- LinearizeNestedList(sub00f, LinearizeDataFrames =
TRUE)
names(sub00f)
# breaks up nest into State/County and State/State

sub00f <- as.data.frame.list(sub00f)
# transforms the flattened/linearized list into a
data.frame
names(sub00f)
head(sub00f, 10)
nrow(sub00f)
# n = 204

# Calc difference btwn measured (AvgPM) and predicted PM
# First using fixed + random effects output
sub00f$D <- (sub00f$AvgPM - sub00f$predictPM.predict.State)

# Summarize this difference
summary(sub00f$D)

# Then just using fixed effects output
sub00f$DFi <- (sub00f$AvgPM -
sub00f$predictPM.predict.fixed)

# Summarize this difference
summary(sub00f$DFi)


############ 2014 ####

# Subset year pm
pm14 <- subset(pm, Year==14)
str(pm14)
# Set buffer as factor
pm14$X10km.buff <- factor(pm14$X10km.buff)
str(pm14)
```

```
# Average PM in 2014
names(pm14)
summary(pm14$AvgPM)

# Final model for pm year (remove year covariate bc only
dealing with one)
MLMfinal14 <- lme (AvgPM ~ X10km.buff + Y_km +
Mean.Elevation..m. + vehstbillmi + enst100 + Season, random
= ~1 | State/County, data = pm14)
summary(MLMfinal14)


### Model validation: Predictive power of model via out-of-
# sample cross validation ###

# Predict 10% without replacement
nrow(pm14)
# PM 14 = 1863
# 10% of 1863 = 186.3
# Predict at 186 sample points
sub14 <- pm14[sample(nrow(pm14), 186),]
nrow(sub14)
head(sub14, 10)

# Predict at the sample points
# Create column for predicted results
sub14$predictPM <- predict(MLMfinal14, sub14, level = 0:1)

# View results
names(sub14)
head(sub14, 10)
tail(sub14, 10)

# "Unnest" county and state
library(devtools)
source_gist(4205477) #loads the function

sub14f <- LinearizeNestedList(sub14, LinearizeDataFrames =
TRUE)
names(sub14f)

sub14f <- LinearizeNestedList(sub14f, LinearizeDataFrames =
TRUE)
names(sub14f)
# breaks up nest into State/County and State/State

sub14f <- as.data.frame.list(sub14f)
```

```
# transforms the flattened/linearized list into a
data.frame
names(sub14f)
head(sub14f, 10)
nrow(sub14f)
# n = 186

# Calc difference btwn measured (AvgPM) and predicted PM
# First using fixed + random effects output
sub14f$D <- (sub14f$AvgPM - sub14f$predictPM.predict.State)

# Summarize this difference
summary(sub14f$D)

# Then just using fixed effects output
sub14f$DFi <- (sub14f$AvgPM -
sub14f$predictPM.predict.fixed)

# Summarize this difference
summary(sub14f$DFi)


################### Joint analysis, 2000 and 2014 ###

# Subset years pm
pmsub <- subset(pm, Year==0 | Year==14)
nrow(pmsub)
# nrow(pmsub) = 3907
nrow(pm14)
# PM 14 = 1863
nrow(pm00)
# PM 00 = 2044

str(pmsub)
# Set buffer as factor
pmsub$X10km.buff <- factor(pmsub$X10km.buff)
str(pmsub)
summary(pmsub$Year)

# Create indicator variable for before (=0, for year 2000
data) and after (=1, for
# year 2014 data) mobile source & power plant standards
# Use fxn ifelse: If year = 14, then I = 1; else I = 0
pmsub$I <- ifelse(pmsub$Year==14, 1, 0)
str(pmsub)

# Set I as factor
pmsub$I <- factor(pmsub$I)
```

```
head(pmsub, 10)
tail(pmsub, 10)

# Final model for joint analysis
# Includes year covariate thru indicator I
# Add in interaction terms:
# For mobile sources, interaction of indicator variable I
# on traffic vehstbillmi
# For power plant, interaction of I on energy (enst100) and
# pwr plant buffer (X10km.buff)

MLMfinalsub <- lme (AvgPM ~ X10km.buff + Y_km +
Mean.Elevation..m. + vehstbillmi + enst100 + Season + I +
I*vehstbillmi + I*enst100 + I*X10km.buff, random = ~1 |
State/County, data = pmsub)
summary(MLMfinalsub)

### Model validation: Predictive power of model via out-of-
# sample cross validation ###

# Predict 10% without replacement
nrow(pmsub)
# 3907
# 10% of 3907 = 390.7
# Predict at 391 sample points
# Function sample(x, size, replace = FALSE, prob = NULL)
subpmsub <- pmsub[sample(nrow(pmsub), 391),]
nrow(subpmsub)

# Predict at the sample points
# Create column for predicted results
subpmsub$predictPM <- predict(MLMfinalsub, subpmsub, level
= 0:1)

# View results
names(subpmsub)
head(subpmsub, 10)

# "Unnest" county and state
# Use LinearizeNestedList function
library(devtools)
source_gist(4205477) #loads the function

subf <- LinearizeNestedList(subpmsub, LinearizeDataFrames =
TRUE)
names(subf)
```

```
subf <- LinearizeNestedList(subf, LinearizeDataFrames =
TRUE)
names(subf)
# breaks up nest into State/County and State/State

subf <- as.data.frame.list(subf)
# transforms the flattened/linearized list into a
data.frame
names(subf)
head(subf, 10)
nrow(subf)
# n = 391

# Calc difference btwn measured (AvgPM) and predicted PM
# using fixed + random effects output
subf$D <- (subf$AvgPM - subf$predictPM.predict.State)

# Summarize this difference
summary(subf$D)
```

Small Scale Trend Analysis

```
# Aims 2 and 3, small scale trend analysis

# The code below is for aim 2, which used the entire NE US
# dataset. For aim 3, the code was applied to a subset of
# the NE US dataset, which included only the PM monitors in
# PA and those within 100 km of the PA border

require(geoR)

Fun1<-function(x)
  return(round(length(x[x>0])/length(x),4))

# Create objects r1 and r2 for means by season and by month

pm00<-read.csv("pm00.csv")
pm14<-read.csv("pm14.csv")

m1<-by(pm00$AvgPM,pm00$Season,mean)
m2<-by(pm14$AvgPM,pm14$Season,mean)
r1<-data.frame(cbind(m1,m2))
names(r1)<-c("Y2000","Y2014")
r1<-rbind(r1,c(mean(pm00$AvgPM),mean(pm14$AvgPM)))
row.names(r1)[5]<-"All"

m1<-by(pm00$AvgPM,pm00$Month,mean)
```

```
m2<-by(pm14$AvgPM,pm14$Month,mean)
r2<-data.frame(cbind(m1,m2))
names(r2)<-c("Y2000","Y2014")
r2<-rbind(r2,c(mean(pm00$AvgPM),mean(pm14$AvgPM)))
row.names(r2)[13]<-"All"

# Kriging based analyses on the near paired design
# This will be done comparing June to June, July to July
# and August to August
# For each analysis we first need to
# (1) establish the coincident monitored locations
# (2) establish what is missing from each year to make a
# complete union of monitored locations
# (3) krige for each year to fill in to make two complete
# paired data sets

#########################################################
June 2000 versus June 2014
#########################################################

pm00June<-pm00[pm00$Month==6,]
pm14June<-pm14[pm14$Month==6,]

# The code below does the coincident locations and finds
# what new locations are needed for each year

n1<-dim(pm00June)[1]
n2<-dim(pm14June)[1]

xy00<-pm00June[,13:14]
xy14<-pm14June[,13:14]
u3<-duplicated(rbind(xy00,xy14))
u4<-rbind(xy00,xy14)[!u3,]

u5<-duplicated(rbind(xy00,u4))
xy00.new<-rbind(xy00,u4)[!u5,]
xy00.new<-xy00.new[-c(1:n1),]

u5<-duplicated(rbind(xy14,u4))
xy14.new<-rbind(xy14,u4)[!u5,]
xy14.new<-xy14.new[-c(1:n2),]

# The combined unioned data set will have n=229
# June 2000: n=170 + 59 new = 229
# June 2014: n=155 + 74 new = 229

# Now we need to get kriged predictions for the 59
# locations in 2000 and 73 locations in 2014
```

```
# Kriging for the 59 locations in 2000

geo00<-as.geodata(cbind(pm00June[,13:14],pm00June[,10]))
v00<-variog(geo00,max.dist=958.6636/2)
plot(v00)
eyefit(v00)

v00.wls<-
variofit(v00,ini.cov.pars=c(56.54,473.34),cov.model="gaussi
an",nugget=2.17)
plot(v00)
lines(v00.wls,lwd=2,col="blue")

krige00<-
krige.conv(geo00,locations=xy00.new,krige=krige.control(obj
.model=v00.wls))

# Kriging for the 74 locations in 2014

geo14<-as.geodata(cbind(pm14June[,13:14],pm14June[,10]))
v14<-variog(geo14,max.dist=958.6636/2)
plot(v14)
eyefit(v14)

v14.wls<-
variofit(v14,ini.cov.pars=c(20.85,485.79),cov.model="gaussi
an",nugget=1.84)
plot(v14)
lines(v14.wls,lwd=2,col="blue")

krige14<-
krige.conv(geo14,locations=xy14.new,krige=krige.control(obj
.model=v14.wls))

# Now combine the observed and predicted for 2000 as well
# as for 2014
# Reorder each file so they are now paired (matched) by
# location
# Generate a PM difference dataframe June 2014 - June 2000

newpm00June<-
rbind(pm00June[,c(13,14,10)],data.frame(xy00.new,AvgPM=krig
e00$predict))

newpm14June<-
rbind(pm14June[,c(13,14,10)],data.frame(xy14.new,AvgPM=krig
e14$predict))
```

```
newpm00June<-
newpm00June[order(newpm00June[,1],newpm00June[,2]),]

newpm14June<-
newpm14June[order(newpm14June[,1],newpm14June[,2]),]

# Check to make sure coordinates are all the same and
# ordered the same between the two files
# Ranges should be 0,0

range(newpm00June[,1]-newpm14June[,1])
range(newpm00June[,2]-newpm14June[,2])

JuneDiffpm<-
data.frame(newpm00June[,1:2],Diff=newpm14June[,3]-
newpm00June[,3])

# Kriging analysis of the paired difference

diff.geo<-
as.geodata(cbind(JuneDiffpm[,1:2],JuneDiffpm[,3]))
vdiff<-variog(diff.geo,max.dist=958.6636/2)
plot(vdiff)
eyefit(vdiff)

vdiff.wls<-
variofit(vdiff,ini.cov.pars=c(21.67,398.6),cov.model="gauss
ian",nugget=3.01)
plot(vdiff)
lines(vdiff.wls,lwd=2,col="blue")

# Below generates kriged predictions (of the differences)
# at the 3920 grid locations
# and also produces 1000 conditionally simulated values at
# each grid location as well.  This is simulating
# from the kriging prediction distribution.  If we average
# the 1000 simulations per grid location we should
# get back (or close to) the kriged prediction.  Instead we
# use this 1000 simulated values to calculate the
# estimated probabilities.

Junekrigediff<-
krige.conv(diff.geo,locations=grid,krige=krige.control(obj.
model=vdiff.wls),

output=output.control(n.predictive=1000))
```

```
sims<-Junekrigediff$simulations

# Map the probabilities and significance

Juneprobs<-apply(sims,1,Fun1)
plot(grid,xlab="",ylab="",axes=F)
points(grid[Junekrigediff$predict<0,1:2],col="blue",pch=16)
points(grid[Juneprobs<.05,1:2],col="yellow",pch=16)
points(diff.geo$coords,pch=16,cex=.75)

probs.geo<-as.geodata(cbind(grid,Juneprobs))
points(probs.geo,pt.divide="data.proportional")


#########################################################
July 2000 versus July 2014
#########################################################

pm00July<-pm00[pm00$Month==7,]
pm14July<-pm14[pm14$Month==7,]

# Coincident locations and what new locations are needed
# for each year

n1<-dim(pm00July)[1]
n2<-dim(pm14July)[1]

xy00<-pm00July[,13:14]
xy14<-pm14July[,13:14]
u3<-duplicated(rbind(xy00,xy14))
u4<-rbind(xy00,xy14)[!u3,]

u5<-duplicated(rbind(xy00,u4))
xy00.new<-rbind(xy00,u4)[!u5,]
xy00.new<-xy00.new[-c(1:n1),]

u5<-duplicated(rbind(xy14,u4))
xy14.new<-rbind(xy14,u4)[!u5,]
xy14.new<-xy14.new[-c(1:n2),]

# The combined unioned data set will have n=231
# July 2000: n=174 + 57 new = 231
# July 2014: n=156 + 75 new = 231

# Kriging for the 57 locations in 2000

geo00<-as.geodata(cbind(pm00July[,13:14],pm00July[,10]))
v00<-variog(geo00,max.dist=958.6636/2)
```

```
plot(v00)
eyefit(v00)

v00.wls<-
variofit(v00,ini.cov.pars=c(38.86,386.14),cov.model="gaussi
an",nugget=0.97)
plot(v00)
lines(v00.wls,lwd=2,col="blue")

krige00<-
krige.conv(geo00,locations=xy00.new,krige=krige.control(obj
.model=v00.wls))

# Kriging for the 75 locations in 2014

geo14<-as.geodata(cbind(pm14July[,13:14],pm14July[,10]))
v14<-variog(geo14,max.dist=958.6636/2)
plot(v14)
eyefit(v14)

v14.wls<-
variofit(v14,ini.cov.pars=c(7.95,572.99),cov.model="gaussia
n",nugget=1.81)
plot(v14)
lines(v14.wls,lwd=2,col="blue")

krige14<-
krige.conv(geo14,locations=xy14.new,krige=krige.control(obj
.model=v14.wls))

# Combine the observed and predicted for 2000 and 2014
# Reorder each file so they are now paired by location
# Generate a PM difference dataframe July 2014 - July 2000

newpm00July<-
rbind(pm00July[,c(13,14,10)],data.frame(xy00.new,AvgPM=krig
e00$predict))

newpm14July<-
rbind(pm14July[,c(13,14,10)],data.frame(xy14.new,AvgPM=krig
e14$predict))

newpm00July<-
newpm00July[order(newpm00July[,1],newpm00July[,2]),]

newpm14July<-
newpm14July[order(newpm14July[,1],newpm14July[,2]),]
```

```
# Make sure coordinates are ordered the same

range(newpm00July[,1]-newpm14July[,1])
range(newpm00July[,2]-newpm14July[,2])

JulyDiffpm<-
data.frame(newpm00July[,1:2],Diff=newpm14July[,3]-
newpm00July[,3])

# Kriging analysis of the paired difference

diff.geo<-
as.geodata(cbind(JulyDiffpm[,1:2],JulyDiffpm[,3]))
vdiff<-variog(diff.geo,max.dist=958.6636/2)
plot(vdiff)
eyefit(vdiff)

vdiff.wls<-
variofit(vdiff,ini.cov.pars=c(18.95,423.51),cov.model="gaus
sian",nugget=2.99)
plot(vdiff)
lines(vdiff.wls,lwd=2,col="blue")

# Kriged predictions of the differences at the 3920 grid
# locations and 1000 conditionally simulated values at each
# grid location
# Use simulated values to calculate estimated probabilities

Julykrigediff<-
krige.conv(diff.geo,locations=grid,krige=krige.control(obj.
model=vdiff.wls),

output=output.control(n.predictive=1000))

sims<-Julykrigediff$simulations

# Map the probabilities and significance

Julyprobs<-apply(sims,1,Fun1)
plot(grid,xlab="",ylab="",axes=F)
points(grid[Julykrigediff$predict<0,1:2],col="blue",pch=16)
points(grid[Julyprobs<.05,1:2],col="yellow",pch=16)
points(grid[Julykrigediff$predict>0,1:2],col="red",pch=16)
points(grid[newprobs>.95,1:2],col="yellow",pch=16)
points(diff.geo$coords,pch=16,cex=.75)

probs.geo<-as.geodata(cbind(grid,Julyprobs))
points(probs.geo,pt.divide="quintiles")
```

```
#########################################################
Aug 2000 versus Aug 2014
#########################################################

pm00Aug<-pm00[pm00$Month==8,]
pm14Aug<-pm14[pm14$Month==8,]

# Coincident locations and what new locations are needed

n1<-dim(pm00Aug)[1]
n2<-dim(pm14Aug)[1]

xy00<-pm00Aug[,13:14]
xy14<-pm14Aug[,13:14]
u3<-duplicated(rbind(xy00,xy14))
u4<-rbind(xy00,xy14)[!u3,]

u5<-duplicated(rbind(xy00,u4))
xy00.new<-rbind(xy00,u4)[!u5,]
xy00.new<-xy00.new[-c(1:n1),]

u5<-duplicated(rbind(xy14,u4))
xy14.new<-rbind(xy14,u4)[!u5,]
xy14.new<-xy14.new[-c(1:n2),]

# The combined unioned data set will have n=232
# Aug 2000: n=175 + 57 new = 232
# Aug 2014: n=156 + 76 new = 232

# Kriging for the 57 locations in 2000

geo00<-as.geodata(cbind(pm00Aug[,13:14],pm00Aug[,10]))
v00<-variog(geo00,max.dist=958.6636/2)
plot(v00)
eyefit(v00)

v00.wls<-
variofit(v00,ini.cov.pars=c(26.83,298.95),cov.model="gaussi
an",nugget=2.37)
plot(v00)
lines(v00.wls,lwd=2,col="blue")

krige00<-
krige.conv(geo00,locations=xy00.new,krige=krige.control(obj
.model=v00.wls))
```

```
# Kriging for the 76 locations in 2014

geo14<-as.geodata(cbind(pm14Aug[,13:14],pm14Aug[,10]))
v14<-variog(geo14,max.dist=958.6636/2)
plot(v14)
eyefit(v14)

v14.wls<-
variofit(v14,ini.cov.pars=c(7.36,398.6),cov.model="gaussian
",nugget=2.24)
plot(v14)
lines(v14.wls,lwd=2,col="blue")

krige14<-
krige.conv(geo14,locations=xy14.new,krige=krige.control(obj
.model=v14.wls))

# Combine observed and predicted for 2000 and 2014
# Reorder each file so matched by location
# Generate a PM difference dataframe Aug 2014 - Aug 2000

newpm00Aug<-
rbind(pm00Aug[,c(13,14,10)],data.frame(xy00.new,AvgPM=krige
00$predict))

newpm14Aug<-
rbind(pm14Aug[,c(13,14,10)],data.frame(xy14.new,AvgPM=krige
14$predict))

newpm00Aug<-
newpm00Aug[order(newpm00Aug[,1],newpm00Aug[,2]),]

newpm14Aug<-
newpm14Aug[order(newpm14Aug[,1],newpm14Aug[,2]),]

# Checking coordinate order

range(newpm00Aug[,1]-newpm14Aug[,1])
range(newpm00Aug[,2]-newpm14Aug[,2])

AugDiffpm<-data.frame(newpm00Aug[,1:2],Diff=newpm14Aug[,3]-
newpm00Aug[,3])

# Kriging analysis of the paired difference

diff.geo<-as.geodata(cbind(AugDiffpm[,1:2],AugDiffpm[,3]))
vdiff<-variog(diff.geo,max.dist=958.6636/2)
```

```
plot(vdiff)
eyefit(vdiff)

vdiff.wls<-
variofit(vdiff,ini.cov.pars=c(9.52,211.76),cov.model="gauss
ian",nugget=2.57)
plot(vdiff)
lines(vdiff.wls,lwd=2,col="blue")

# Kriged predictions of the differences at the 3920 grid
# locations and 1000 conditionally simulated values at each
# grid location
# Use simulated values to calculate estimated probabilities

Augkrigediff<-
krige.conv(diff.geo,locations=grid,krige=krige.control(obj.
model=vdiff.wls),

output=output.control(n.predictive=1000))

sims<-Augkrigediff$simulations

# Map the probabilities and significance

Augprobs<-apply(sims,1,Fun1)
plot(grid,xlab="",ylab="",axes=F)
points(grid[Augkrigediff$predict<0,1:2],col="blue",pch=16)
points(grid[Augprobs<.05,1:2],col="yellow",pch=16)
points(grid[Augkrigediff$predict>0,1:2],col="red",pch=16)
points(grid[newprobs>.95,1:2],col="yellow",pch=16)
points(diff.geo$coords,pch=16,cex=.75)

probs.geo<-as.geodata(cbind(grid,Augprobs))
points(probs.geo,pt.divide="quintiles")
```

Curriculum Vitae

Stacy E. Woods
3147 Tilden Drive, Baltimore, MD 21211
(954) 288 - 5112

A creative public health professional with biostatisical, environmental health, epidemiology, and risk assessment experience dedicated to public policy to improve environmental health.

EDUCATION

Doctor of Philosophy, Environmental Health Sciences Expected June 2016
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Risk Sciences and Public Policy Certificate (2013)
Dissertation: "Investigating the space-time variation in fine particulate matter pollution in the Northeastern United States, 2000 – 2014"
Advisor: Frank Curriero, PhD

Master of Public Health December 2010
Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
Capstone: "Spatial Analysis of Lyme Disease in Howard County, Maryland"

Bachelor of Science in Entomology and Nematology December 2001
University of Florida, Gainesville, FL

WORK HISTORY

- **Mirzayan Science and Technology Policy Fellow**, The National Academy of Sciences, Washington, DC
- **Senior Research Assistant,** Center to Reduce Cancer Disparities, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
- **Public Health Intern**, Howard County Health Department, Columbia, MD
- **Biologist**, University of Florida, Florida Medical Entomology Laboratory, Vero Beach, FL

PUBLICATIONS

- Jennings JM, **Woods SE**, Curriero FC (2013). The spatial and temporal association of neighborhood drug markets and rates of sexually transmitted infections in an urban setting. Health & Place Volume 23, September 2013, Pages 128–137.

PUBLICATIONS IN PREPARATION

- **Woods, SE**, Waugh, DW, Koehler, KA, Davis, MF, Fox, MA, Rule, AM, and Curriero, FC (2016). The characterization of $PM_{2.5}$ in the northeastern United States as a function of environmental determinants. [In Progress]
- **Woods, SE**, Waugh, DW, Koehler, KA, Davis, MF, Fox, MA, Rule, AM, and Curriero, FC (2016). Investigating the large scale trends and small scale spatial variation in $PM_{2.5}$ pollution and the efficacy of federal emissions regulations in reducing fine particulate pollution in the northeastern United States. [In Progress]
- **Woods, SE**, Waugh, DW, Koehler, KA, Davis, MF, Fox, MA, Rule, AM, and Curriero, FC (2016). The association of the fracking industry with small scale variability in $PM_{2.5}$ pollution in Pennsylvania, 2004 - 2014. [In Progress]

PRESENTATIONS

- "Improved air quality from fuel standards." Presentation at the Energy & Health in Maryland meeting, Maryland Environmental Health Network. November 20, 2014.
- "Investigating the local air quality impact of a new national gasoline rule." Johns Hopkins Bloomberg School of Public Health. 8 October 2013.
- "GIS and Spatial Analysis in Public Health." Safe Kids Worldwide Childhood Injury Prevention Conference. 12 June 2013.
- "Spatial analysis of Lyme disease incidence in Howard County, Maryland." Poster presentation, American Public Health Association Annual Meeting and Exposition. 1 November 2011.
- "Lyme Disease Clustering in Howard County, Maryland." PHASE Symposium, Maryland Department of Health and Mental Hygiene. 14 May 2010.

PROFESSIONAL MEMBERSHIPS

- American Public Health Association
- Maryland Environmental Health Network

HONORS AND AWARDS

- **C. Sylvia and Eddie C. Brown Scholar in Community Health**, Johns Hopkins Bloomberg School of Public Health
- **Dr. C. W. Kruse Memorial Fund Award**, Department of Environmental Health Sciences, Johns Hopkins Bloomberg School of Public Health
- **Delta Omega Honorary Society Alpha Chapter**, Johns Hopkins Bloomberg School of Public Health (Nominated 2016)